DOCUMENT RESUME

ED 092 601 95 TM 003 784

AUTHOR	Bianchini, John C.; Loret, Peter G.
TITLE	Anchor Test Study. Final Report. Project Report.
INSTITUTION	Educational Testing Service, Berkeley, Calif.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	[74]
CONTRACT	OEC-0-71-4758(248)
NOTE	295p.; For related documents, see TM003785-817
EDRS PRICE	MF-$0.75 HC-$13.80 PLUS POSTAGE
DESCRIPTORS	*Administration; Comparative Statistics; Correlation; Data Collection; *Elementary School Students; *Equated Scores; Grade 4; Grade 5; Grade 6; Measurement Techniques; Norms; Raw Scores; Reading; Reading Comprehension; *Reading Tests; *Research Design; Sampling; Social Differences; Standardized Tests; Vocabulary
IDENTIFIERS	*Anchor Test Study; California Achievement Tests; Comprehensive Tests of Basic Skills; Gates MacGinitie Test; Iowa Test of Basic Skills; Metropolitan Achievement Tests; Sequential Tests of Educational Progress; SRA Achievement Series; Stanford Achievement Tests

ABSTRACT

The Anchor Test Study provides a method for translating a pupil's score on any one of eight widely used standardized reading tests for Grades 4, 5, and 6 to a corresponding score of any of the other seven tests, as well as furnishing new nationally representative norms for each of the eight tests. In addition, the Study presents new estimates of alternate form reliability for each test, provides estimates of the intercorrelations among the tests, and explores empirically some methodological questions in test-equating. Contained in the Project Report is descriptive information about the Anchor Test Study (ATS), administration and processing, the sample design for ATS, methodology including restandardization and equating phases, results, recommendations, and appendices. The tests used in the study are: California Achievement Tests (1970), Comprehensive Tests of Basic Skills (1968), Iowa Test of Basic Skills (1971), Metropolitan Achievement Tests (1970), Sequential Tests of Educational Progress (1969), SRA Achievement Series (1971), and Stanford Achievement Tests (1964). Equating and norming data on the eighth test in the Study, the Gates MacGinitie (1964), are included in volumes 31, 32, and 33 of the Anchor Test Study. (RC)

ED 092601

anchor test study

FINAL REPORT

Project Director: Peter G. Loret

Project Coordinator: Alan Seder

Principal Investigator: John C. Bianchini

Associate Investigator: Carol A. Vale

With Sampling Contributions by:

David W. Chapman
Morris H. Hansen
Thomas W. McKenna

Technical Advisors: William H. Angoff
Richard S. Levine
Frederic M. Lord

TM 003 784

December 1972

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

This report was prepared under Contract No. OEC-0-71-4758(248) and submitted to the National Center for Educational Statistics, Office of Education, U.S. Department of Health, Education, and Welfare. The contractor was encouraged to exercise professional judgment in its contents; therefore, this report does not necessarily reflect positions and policies of the Government. In particular, no official endorsement of the tests mentioned in the report should be inferred.

ACKNOWLEDGEMENT

Many individuals contribute to an undertaking of this magnitude. The administration, staff, and pupils of the cooperating states, districts/dioceses, and schools are certainly the most important. Although a listing of all these individuals would be impossible, the names of the participating schools, districts, and states are listed in Volume XXVIII of this report. Their cooperation and assistance are gratefully acknowledged.

Important contributions were made by the Educational Testing Service staff in the Berkeley, California office. Several individuals were specifically concerned with the planning and execution of the entire study: Alan Seder shared with me the tasks of general planning and conduct of the study, with particular emphasis on the invitational, data collection, and scoring/reporting phases; John C. Bianchini and Carol N. Vale, members of the ETS Research staff, planned and executed all phases of the project data analysis. Several senior ETS staff members, particularly William H. Angoff, Richard S. Levine, and Frederic M. Lord, provided valuable advice on methodology. Operational assistance was capably provided by Richard C. Minard, Roger W. Ervin, and Chris Rivers. This 34-volume Final Report is indeed weighty testimony to the magnitude of the now-completed task.

The subcontractor for the sampling portion of this study was Westat Research, Inc. Morris H. Hansen, Thomas W. McKenna, and David W. Chapman of the Westat staff contributed significantly to the study design.

We wish to thank the members of the Anchor Test Study Advisory Council who served as consultants for many aspects of the Study. In addition, our thanks are extended to Richard M. Jaeger (formerly of the U.S. Office of Education) who originally formulated the Study and saw it through its initial phases, and Charles H. Hammer, who served as USOE Project Officer, for the Study.

Peter G. Loret
Project Director

TABLE OF CONTENTS

	Page
CHAPTER 1 THE ANCHOR TEST STUDY	
1. The Nature of the Study	1
2. The Need for the Study	2
3. The Focus of the Study	3
4. Tests Included	3
5. Design and Conduct of the Study	5
CHAPTER 2 ADMINISTRATION AND PROCESSING	
1. The Advisory Council	6
2. The Invitational Phase	8
2.1 State Invitations	8
2.2 District Invitations	11
2.3 School Invitations	13
2.4 Participation	15
3. Materials Preparation and Shipping	17
3.1 Packing, Shipping, and Control	18
4. Materials Receipt/Scoring and Reporting	23
4.1 Receiving	24
4.2 Scoring and Reporting	25
CHAPTER 3 SAMPLE DESIGN FOR THE ANCHOR TEST STUDY	
1. Introduction	27
2. Brief Summary of the Sample Designs	32
2.1 The Sample Design for the Restandardization Study	32
2.2 The Sample Design for the Equating Study	35
3. Some of the Decisions, Analyses, and Considerations that Influenced the Sample Design	39
3.1 The Universe of Schools and the Sources Used to Construct the Sampling Frame	39
3.2 The Use of Schools as Sampling Units	40
3.3 The Use of Mail and Telephone Communication with Schools and School Districts	43
3.4 The Inclusion in the Study of All Eligible Pupils in Each Participating School	43
3.5 The Approximate Optimum Allocation of the Restandardization Sample to Enrollment Size Classes	44
3.6 The Use of Measures of Size to Allocate the School Sample to Enrollment Size Classes	48
3.7 The Inclusion in the Sample of All of the Pupils Contained in Grades 4, 5, and 6 in each Participating School	51
3.8 The Possible Use of School District Information from the 1970 Census to Improve Sample Design	52
4. The Stratification Variables Used in the Study	53
4.1 Introduction	53
4.2 Type of Control (Public or Private)	54
4.3 Size of School District and Population of County	54
4.4 Income Class	55

4.5 Percent Minority 58
4.6 Geographic Region 59
4.7 Degree of Urbanization 60
5. Stratification and Selection Procedures for the Restandardization Study 61
5.1 Basic Stratification and Selection Design 61
5.2 The Definition of the 47 Major Strata 63
5.3 The Definition of Final Strata 65
5.4 Selection of the Primary Sample for the Restandardization Study 70
5.5 The Selection and Use of Substitutes for the Primary Restandardization Sample 73
6. Stratification and Selection Procedures for the Equating Study 76
6.1 The Definition of Pseudo Schools 76
6.2 The Definition of the 27 Major Strata 78
6.3 The Definition of Final Strata and Selection of the Primary Sample 82
6.4 The Selection and Use of Substitutes for the Primary Equating Sample 84
7. Selection of a Supplemental Sample of Public Schools Omitted from the Sampling Frame for the Restandardization Study 86
7.1 Summary of the Sample Supplementation Procedure 86
7.2 The Selection of a Sample of 178 Public School Districts for Sample Supplementation 88
7.3 The Selection of the Sample of Missed Schools from the Districts Selected for Supplementation 93
8. The Assignment of Tests to Schools for the Equating Study ... 96
8.1 The Possible Test Assignments 96
8.2 The Assignment of the Schools in the Primary Equating Sample to Test Assignment Matrices 99
8.3 The Assignment of Classes to Test Sequences 102
8.4 Resolving Conflicts of Some of the Test Assignments 105
9. The Assignment of Weights to the Respondents 107
9.1 Summary of the Weighting Procedures 107
9.2 School Weights 108
9.3 Pupil Weights for the Restandardization Study 111
9.4 Pupil Weights for the Equating Study 114
10. Survey Estimates 116
10.1 Introduction 116
10.2 Estimates of Pupil Percentiles and Percentile Ranks 117
10.3 The Estimate of the Pupil Mean 119
10.4 Estimation of School Percentiles and Percentile Ranks .. 120
11. Variance Estimation 121
11.1 Brief Summary of the Variance Estimation Procedure for the Restandardization Study 121
11.2 Estimation of the Variances for the Equating Study 123

CHAPTER 4 METHODOLOGY

1. Restandardization Study 128
1.1 Weighting Procedure 128

1.2 MAT Individual Norms 130
1.3 Individual Norms for CAT, CTBS, ITBS, STEP II, SRA, SAT 131
1.4 School Mean Norms for the MAT 132
1.5 School Mean Norms for CAT, CTBS, ITBS, STEP II, SRA, SAT 133
2. Equating Study 134
2.1 Weighting Procedure 140
2.2 Notation 141
2.3 Equipercentile Method: Procedure 1 142
2.4 Equipercentile Method: Procedure 2 145
2.5 Equipercentile Method: Average of Procedures 1 and 2 .. 145
2.6 Equipercentile Method: Procedure 3 146
2.7 Linear Method: Procedure 1 147
2.8 Linear Method: Procedure 2 155
2.9 Linear Method: Average of Procedures 1 and 2 155
2.10 Linear Method: Procedure 3 156
2.11 Estimation of Equating Error 157
2.12 Conditional Error of Equating 158
2.13 Equating Analyses for Special Subgroups 160
3. Correlations 162

CHAPTER 5 RESULTS

1. Restandardization Study 164
1.1 Restandardization Sample 164
1.2 Individual and School Mean Norms 166
1.3 Comparisons Between Anchor Test Study and Test Publishers' Norms 171
2. Equating Study 174
2.1 Equating Sample 174
2.2 Comparability of the Tests 175
2.3 Equating Results for the Total Sample 181
2.4 Evaluation of Equating Procedures and Methods 187
2.5 Equating Error 188
2.6 Subgroup Equating Results 192

CHAPTER 6 RECOMMENDATIONS

1. Selection of the Anchor Test 197
2. Changes in Equating Procedures and Methods 197
3. Generalizability of Equating Results to Special Subgroups ... 198
4. Operational and Administrative Recommendations 199

APPENDICES

A. The Determination of Approximately Optimum Probabilities of Selection by Size Class of School 203
B. The Determination of the Set of Measures of Size 224
C. The Formation of Major Strata 228
D. Approximation of District Selection Probabilities 238
E. Survey Estimates and Corresponding Variance Estimates for the Restandardization Study 241
F. Classification Procedure for "Weighting Classes" and Definitions of Eligibility and Usable Cases 259

G. Tukey-Cureton Smoothing Procedure 261
H. Notation 262
J. Frequency Estimation Procedure 269
K. Equipercentile Equating Procedure 275
L. Computational Formulas 278

LIST OF TABLES

3-1 Measures of Size 62

3-2 Major and Final Stratum Sizes for the Restandardization Study . 67

3-3 Major and Final Stratum Sizes for the Equating Study 83

3-4 District Subsampling Fractions and Selection Probabilities 90

3-5 The Test Assignment Matrix 98

3-6 The Grouping of Final Strata into Principal Strata 100

3-7 Testing Sequence Assignment Patterns 104

3-8 The 8 x 8 Orthogonal Matrix Used to Define the Set of Eight Balanced Half Samples 125

4-1 Range of Percentile Ranks Defining the Stanine Scale 130

4-2 Reference Table for Individual Score Norms for the Non-MAT Tests 132

4-3 Reference Table for School Mean Norms for the Non-MAT Tests ... 134

4-4 Design for Equating Study Test Administration 135

4-5 General Design for Equating Study 136

4-6 Reference Table for Equipercentile Method/Procedure 1 Equating Results 144

4-7 Reference Table for Equipercentile Method/Procedure 2 Equating Results 145

4-8 Reference Table for Equipercentile Method/Average of Procedures 1 and 2 Equating Results 146

4-9 Reference Table for Equipercentile Method/Procedure 3 Equating Results 147

4-10 Reference Table for Linear Method/Procedure 1 Equating Results 154

4-11 Reference Table for Linear Method/Procedure 2 Equating Results 155

4-12 Reference Table for Linear Method/Average of Procedures 1 and 2 Equating Results 156

4-13 Reference Table for Linear Method/Procedure 3 Equating Results 156

4-14 Reference Table for Estimates of Equating Error 158

4-15 Reference Table for Correlation Matrices 163

5-1 Restandardization Sample: Numbers and Percentages of Schools by Major Stratum 165

5-2 Restandardization Sample: Numbers and Percentages of Public Schools by States 167

5-3 Restandardization Sample: Numbers and Percentages of Catholic Schools by States 168

5-4 Restandardization Sample: Numbers and Percentages of Private Non Catholic Schools by States 169

5-5 Restandardization Sample: Number of Pupils by Racial Groups and IQ Groups with Usable or Absentee, Incomplete, or Invalid Answer Sheets 170

5-6 Equating Sample: Numbers and Percentages of Schools by Major Stratum 176

5-7 Equating Sample: Numbers and Percentages of Public Schools by States 177

5-8 Equating Sample: Numbers and Percentages of Catholic Schools by States 178

5-9 Equating Sample: Numbers and Percentages of Non Catholic Schools by States 179

5-10 Equating Sample: Number of Pupils by Racial Groups and IQ Groups with Usable or Absentee, Incomplete, or Invalid Answer Sheets.. 180

5-11 Distribution of Correlations Corrected for Attenuation 182

5-12 Reliabilities and Ranges of Averaged Intercorrelations Among Tests (Grade 4) 183

5-13 Reliabilities and Ranges of Averaged Intercorrelations Among Tests (Grade 5) 184

5-14 Reliabilities and Ranges of Averaged Intercorrelations Among Tests (Grade 6) 185

5-15 Equating Study (Grade 4): Estimates of Averaged Error in Equating the Reading Tests to the MAT (70 ED.) Elementary, Form F (Average of Procedures 1 and 2, Equipercentile) 189

5-16 Equating Study (Grade 5): Estimates of Averaged Error in Equating the Reading Tests to the MAT (70 ED.) Intermediate, Form F (Average of Procedures 1 and 2, Equipercentile) 190

5-17 Equating Study (Grade 6: Estimates of Averaged Error in Equating the Reading Tests to the MAT (70 ED.) Intermediate, Form F (Average of Procedures 1 and 2, Equipercentile) 191

CONTENTS OF VOLUMES I - XXXIII

Volume	
I	Individual Norms and School Mean Norms, Grades 4, 5, 6
II	Equating Tables, Error of Equating, and Correlations, Grade 4
III	Equating Tables, Error of Equating, and Correlations, Grade 5
IV	Equating Tables, Error of Equating, and Correlations, Grade 6
V	Equating Procedure Comparison Graphs, Grade 4
VI	Equating Method Comparison Graphs, Grade 4
VII	Equating Procedure Comparison Graphs, Grade 5
VIII	Equating Method Comparison Graphs, Grade 5
IX	Equating Procedure Comparison Graphs, Grade 6
X	Equating Method Comparison Graphs, Grade 6
XI	Subgroup Equating Tables - Boys, Grades 4, 5, 6
XII	Subgroup Equating Tables - Girls, Grades 4, 5, 6
XIII	Subgroup Equating Tables - High IQ Children, Grades 4, 5, 6
XIV	Subgroup Equating Tables - Medium IQ Children, Grades 4, 5, 6
XV	Subgroup Equating Tables - Low IQ Children, Grades 4, 5, 6
XVI	Subgroup Equating Tables - Caucasian Children, Grades 4, 5, 6
XVII	Subgroup Equating Tables - Black Children, Grades 4, 5, 6
XVIII	Subgroup Equating Tables - Spanish-Surnamed Children, Grades 4, 5, 6
XIX	Subgroup Equating Tables - High SES Children, Grades 4, 5, 6
XX	Subgroup Equating Tables - Medium SES Children, Grades 4, 5, 6
XXI	Subgroup Equating Tables - Low SES Children, Grades 4, 5, 6
XXII	Subgroup Comparison Graphs - Linear Equating Method, Grade 4
XXIII	Subgroup Comparison Graphs - Equipercentile Equating Method, Grade 4
XXIV	Subgroup Comparison Graphs - Linear Equating Method, Grade 5
XXV	Subgroup Comparison Graphs - Equipercentile Equating Method, Grade 5
XXVI	Subgroup Comparison Graphs - Linear Equating Method, Grade 6
XXVII	Subgroup Comparison Graphs - Equipercentile Equating Method, Grade 6
XXVIII	Pretest Information Form
XXIX	Ancillary Data Obtained on Individual Pupils
XXX	Norms Comparisons Conditional Error of Equating Quality Control Results Convergence of Equating Iterations
XXXI	Project Report
XXXII	Equating Tables - Grades 4, 5, 6
XXXIII	Error of Equating Correlations Conditional Error of Equating Procedure Comparison Graphs Individual Norms and School Mean Norms Norms Comparisons Grades 4, 5, and 6

CHAPTER 1

THE ANCHOR TEST STUDY

1. The Nature of the Study

The Anchor Test Study had two major objectives: to provide a method by which one may translate a child's score on any one of seven widely used standarized reading tests into a score on any of the other tests, and to provide new nationally representative norms for each of these seven tests. Tables for this purpose have been prepared for tests appropriate for 4th grade, 5th grade, and 6th grade children.

Children tested in this Study were selected, using the latest principles of random sampling, from data already available to the U.S. Office of Education.

The Equating Study involved the administration of pairs of tests to 4th, 5th, and 6th graders in a sample of public and nonpublic schools. Two appropriate reading tests were administered to each participating student to derive equivalent scores on the seven tests.

The Restandardization Study involved the administration of the Metropolitan Reading Test at Grades 4, 5, and 6, in a randomly selected, stratified sample of public and nonpublic schools. The Word Knowledge and Reading subtests of the Metropolitan Reading Test were administered to the participants. New norms for the Metropolitan Reading Test have been constructed from the results of this administration. The results of the Equating Study and those of the Restandardization Study were combined to produce new norms for the other six reading tests.

2. The Need for the Study

For more than 30 years, those concerned with educational testing and evaluation have recognized that standardized tests, which claim to measure the same skills, actually vary considerably in their approach to measurement. Thus, different publisher's tests of "reading" may differ in content, scales of measurement, and the design of their norming studies. Ordinarily, these differences might be regarded as the result of normal and expected variation in measurement philosophy and opinion among members of the testing profession, and accepted simply as an area of academic disagreement. However, as early as 1941, it was recognized that some of the differences were legitimately avoidable and that reduction of differences would be of major advantage to educational practitioners and researchers.

Differences in test content are unavoidable unless we accept only one "right way" to measure specific skills. Differences in norming and standardization can, however, be avoided.

Prior to the completion of the Anchor Test Study, schools accustomed to one test could not move to another test without breaking the continuity of their testing program. Schools within the same district could not merge or compare their data, nor could they participate in a mutually beneficial evaluation of experimental programs, unless they all used the same test. These problems were not restricted to schools. What was perhaps of greater concern were the problems faced by central agencies, such as those at the county, state, and federal level; these agencies could not merge test results unless all schools involved in the comparison used the same test.

The Anchor Test Study permits the merging of data on different reading tests and thus eliminates many of these problems. School counselors can convert the test scores of students who transfer into their districts to

scores on the reading tests they normally administer. School system offices can convert to a single test, scores on different tests used by different schools, and thus compile reading achievement statistics for the entire district. State departments of education can allow freedom in school districts' choices of reading tests and will still be able to compile statewide reports on reading achievement.

3. The Focus of the Study

The ability to read is fundamental to accomplishment in most academic subjects. The choice of upper primary grades for the Study was made because reading is so crucial early in the child's educational career. Also, available data on testing showed that school systems throughout the nation administer reading tests most frequently in the upper elementary grades.

Equivalency tables based on truly representative national samples were needed by many school systems, state departments of education, and at the federal level. These educational agencies additionally needed national norms which included appropriate representation of inner city children, nonpublic school children, and other groups that have been generally under-represented in previously available national norms.

4. Tests Included

The standardized tests included in the Study were among those most commonly used by school systems throughout the nation.

The editions of tests used in this Study were consistently the latest available from the publishers as of Spring 1972. The results of the Study should, therefore, be useful for many years to come. The levels of tests used

were those recommended by the test publishers most often for children in grades 4, 5, and 6. The test editions, forms, and levels used are listed below.

Title/Edition/Form	Publisher	Subtests	LEVEL USED FOR GRADE: 4	5	6
California Achievement Tests (1970)-Reading, Forms A and B	CTB/ McGraw-Hill	Reading Vocabulary Reading Comprehension	Level 3	Level 3	Level 4
Comprehensive Tests of Basic Skills (1968), Forms Q & R	CTB/ McGraw-Hill	Reading Vocabulary Reading Comprehension	Level 2	Level 2	Level 3
Iowa Test of Basic Skills (1970), Forms 5 and 6	Houghton Mifflin	Vocabulary Reading Comprehension	Level 10	Level 11	Level 12
Metropolitan Reading Tests (1970), Forms F and G	Harcourt Brace Jovanovich	Word Analysis Reading	Elementary	Intermediate	Intermediate
Sequential Tests of Educational Progress, STEP Series II (1969), Forms A and B	Educational Testing Service	Reading	Level 4	Level 4	Level 4
SRA Achievement Series (1971), Forms E and F	Science Research Associates	Vocabulary Reading	Blue Edition	Blue Edition	Green Edition
Stanford Reading Tests (1964), Forms W and X	Harcourt Brace Jovanovich	Word Meaning Paragraph Meaning	Intermediate I	Intermediate II	Intermediate II

5. Design and Conduct of the Study

The U.S. Office of Education developed the basic plan and detailed specifications for the Anchor Test Study. The Study was conceived as an integral part of the Office of Education's program of evaluation in elementary and secondary education. It was apparent from the outset that the results of the Study would be useful at all levels of educational administration: evaluative results would be useful to states and the Federal government, and local school systems could enjoy increased flexibility in achievement testing.

The contract for conduct of the Anchor Test Study was awarded to Educational Testing Service (ETS), a nonprofit, nonstock corporation. The Study was conducted by ETS's Western Office in Berkeley, California, with extensive use of the technical expertise of the statistical and psychometric staff in its Princeton, New Jersey office.

CHAPTER 2

ADMINISTRATION AND PROCESSING

1. The Advisory Council

Under the provisions of the Anchor Test Study contract, an Advisory Council was assembled to consult with the contractor (Educational Testing Service) and the U.S. Office of Education (USOE) on matters of technical design and administrative practice. A list of fifteen Council nominees, alternates, and a chairman was prepared jointly by Educational Testing Service and the USOE to represent the following groups:

The Council of Great Cities Schools
The International Reading Association
The National Association of Elementary School Principals
The National Council on Measurement in Education
The National School Board Association

Chief State School Officers
Large City School Superintendents
State Testing Directors
Title I State Coordinators

Curriculum
Minority group representatives
  Spanish-surnamed
  Black
Nonpublic education

A letter of invitation was prepared by Educational Testing Service, approved by the USOE, and released to all nominees on August 17, 1971.

During the next several weeks, telephone follow-ups were conducted to determine the status of outstanding invitations. Details of the first scheduled meeting (in Princeton, September 10, 1971) were sent to those individuals agreeing to serve on the Advisory Council.

Although there was only a three-week interval between the initial invitations and the first scheduled meeting, approximately one-half the members were able to attend. Subsequent meetings were held on November 12th (Princeton), February 4, 1972 (Berkeley, California), and September 15, 1972 (Princeton). A complete list of Advisory Council members may be found in Volume XXVIII.

Summaries of the meetings were prepared by ETS and distributed to all Council members and the USOE. The first meeting dealt with an introduction to the project and a review of the school invitation process. The second meeting concerned itself with additional aspects of the invitation phase and the reporting phase. The third meeting included a review of the project status, exploration of additional areas for project data analyses, and plans for the dissemination of the results. The final meeting included a review of the Study and an examination of the preliminary findings.

2. The Invitational Phase

The invitational phase of the Anchor Test required invitation letters addressed to the responsible individuals at three levels: Chief State School Officer, District (Local Education Agency) Superintendent, and School Principal. Draft letters were reviewed and approved by the U.S. Office of Education (USOE) prior to the preparation of final copy.

Specifications required that approval be received from a higher administrative level prior to the issuance of an invitation to any potential participant, i.e., it was necessary to obtain approval from the State's Chief School Officer before issuing any District (LEA) Superintendent invitation. Similarly, it was necessary to receive District Superintendent approval prior to the issuance of school invitations.

2.1 State Invitations

During the second week of October 1971, fifty-two copies of the Chief State School Officer invitation letters were received from USOE. After the text had been approved by USOE, the letters were prepared by ETS, and final copies transmitted to USOE for signature. The letters were signed by S.P. Marland, Jr., U.S. Commissioner of Education, and forwarded to Florida for the co-signature of Floyd T. Christian, President, Council of Chief State School Officers. ETS was then prepared to begin the invitational phase by releasing this initial mailing to the Chief State School Officers in the fifty states, the District of Columbia, and the National Catholic Education Association.

The Chief State School Officer (CSSO) invitation described the Study, its background, and the benefits to be gained from its success. In

addition to the actual invitation letter, the mailing included a reply form, a model endorsement letter and a business reply envelope. The CSSO was asked to approve the participation of his/her state in the Anchor Test Study, to prepare a letter of endorsement for District Superintendents in the state, and to appoint a State Coordinator for the project. The State Coordinator's function was to 1) assure the existence of selected schools, 2) provide the names of District Superintendents in the selected school districts, 3) assist in the resolution of any communication problems in the state, and 4) provide liason between ETS and the Chief State School Officer.

Forty-nine of the fifty Chief State School Officers agreed to support and endorse the Anchor Test Study. Endorsements were also received from the executive education officers in the District of Columbia and the National Catholic Education Association. Responses were received between October 18 and December 21, 1971 from all states except one. However, that State indicated that District Superintendents could be directly invited to participate in the project. The lack of this State endorsement seemed to have little effect on the participation of schools and districts.

In agreeing to participate in the project, the Chief State School Officers were given two alternate methods for endorsing the Study. One alternative was the ETS duplication and distribution of an endorsement letter prepared by the Chief State School Officer. This letter would be enclosed with the District invitation. The other alternative was the preparation of an endorsement letter which the Chief State School Officer would distribute directly to the selected Districts. A model endorsement

letter was provided for the Chief State School Officer's consideration. Twenty-six participants chose to provide ETS with an endorsement letter. Twenty-five Chief State School Officers chose to send their endorsement letters directly to the selected Districts.

As endorsements were received from each Chief State School Officer, it was possible to contact the person designated as State Coordinator A letter acknowledging this individual's appointment was prepared and sent to him/her, together with a computer-prepared listing of the selected Districts in the state. Twenty-five copies of a brochure entitled "A Description of the Anchor Test Study" were also enclosed, in order to provide information for all those who needed to be made aware of the Anchor Test Study. This was the first use of the brochure; the schedules for writing, design, and UBOL review did not permit inclusion in the Chief State School Officer's mailing, as originally planned. Two alternate cover letters, depending on which endorsement method had been selected by the Chief State School Officer, were prepared for release to the Coordinators. When the Chief State School Officer had selected direct endorsement release from his office, the State Coordinator was reminded to proceed with the immediate release of these endorsements. All State Coordinators were asked to supply the names and addresses of the District Superintendents whose schools had been selected for participation in either the primary or back-up samples. In many cases, the Coordinators supplied state directories that listed all District Superintendents, while in others, the Coordinator only supplied information on sample selectees. Contacts with State Coordinators were made during the latter part of October and continued through the month of December, as agreements were received from the Chief State School Officers. State Coordinators were generally very prompt in

supplying the required District information.

2.2 District Invitations

Upon receipt of the District Superintendents' names and addresses, it was possible to begin the district invitation process. Since the beginning of this phase depended on replies from the State Coordinators, the relase of district invitations varied from state to state, beginning in November 1971 and lasting through January 1972.

District invitations consisted of four components: a letter of invitation, a list of selected primary and back-up schools within the district, four brochures, and a reply postcard. The letter of invitation was reviewed and approved by USOE and signed by Richard M. Jaeger, Acting Director, Federal/State Developmental Staff. Different versions of the invitation letters were designed for: 1) a District invitation, with a copy of the endorsement letter provided by the Chief State School Officer enclosed, 2) a District invitation referring to a special Anchor Test Study endorsement letter being sent directly by the Chief State School Officer, 3) a District invitation making no reference to endorsement by the Chief State School Officer, 4) invitations specific to Catholic dioceses rather than public school personnel, and 5) special invitations to the fifty largest school districts, including a separate detailed memorandum on sampling methodology.

District Superintendents were asked to approve the participation of selected schools in their districts and to appoint a District Coordinator. The District Coordindator would verify the existence of selected schools (as well as the existence of grades 4, 5, and 6), provide the names and addresses of school principals, help with the distribution of school

2.3 School Invitations

With the return of the District Superintendent's reply card, preparations were begun for the mailing of individual school invitation packets. If the Superintendent supplied the names of individual school principals, an invitation packet was prepared based on that information. If not, the District Coordinator was contacted directly and the names and addresses of the selected school principals were requested. All school invitation packets, although addressed to individual principals, were mailed to the responsible District Coordinator. This permitted direct distribution through the District office, and school principals immediately became aware that the District had approved participation. It also served to introduce the individual appointed as District Coordinator. A cover letter, sent with these packets, requested that the District Coordinator distribute the invitations to the schools, help in describing the project to the principals, and assure that the school(s) completed and returned the reply forms contained in the invitation packet.

Several variations of the school invitation letter were prepared by Educational Testing Service and approved by USOE. All versions of the letter were released under the signature of Richard M. Jaeger, Acting Director, Federal/State Developmental Staff. Specific letters of invitation were prepared for Norming schools, Equating schools, private and Massachusetts Norming schools, and private and Massachusetts Equating schools. The school invitation provided the initial contact with the selected private schools.

The invitation letter requested that the school principal approve participation for his/her school and appoint an individual to serve as School Coordinator. This individual would have primary responsibility for the actual test administration and data collection. The School

Coordinator would complete a Pre-Test Information Form, be responsible for proper test administration, receive and distribute test materials, and complete the school and class summary forms. Since these tasks would involve other than class time, the School Coordinator would receive an honorarium based on the number of pupils tested:

Honorarium for Test Coordination

149 or fewer pupils - $30

150 to 249 pupils - $40

250 or more pupils - $50

In addition to the letter of invitation, selected schools received four items in their invitation packet. A copy of the brochure entitled "A Description of the Anchor Test Study" was enclosed. A reply postcard was provided for the name, address, and phone number of the School Coordinator. The card also provided for indications that the school was still considering the request and/or wished to have additional information. The postcard was completed by the principal and given to the School Coordinator with the remaining two items in the packet, the Pre-Test Information Form and a business reply envelope. Agreement to participate was indicated by the return of the completed postcard and the Pre-Test Information Form.

School invitation packets were first released in December and mailings continued until April. Schools unable to accept the invitation were replaced by schools designated as back-ups in the sample. Depending upon the characteristics of the originally selected school, the specific back-ups were located either in the same district, or in another district and/or state. In some cases, this required going back one step in the process and issuing an invitation to a District Superintendent not

previously invited.

During the entire school invitation process, the District Coordinators, State Coordinators, and the USOE were kept informed of the status of all school invitations. District Coordinators were provided with a weekly listing of schools agreeing to participate, schools unable to accept, those invitations still outstanding, and back-up schools selected but not yet invited. The names of School Coordinators were also provided. Statewide summaries were provided for each State Coordinator, and a project summary was provided to the USOE. A cover letter requested the District Coordinator's assistance in following up non-responding schools. In addition, telephone follow-ups with District Coordinators and principals were made to elicit rapid responses. In many cases, School Coordinator assignments were made by the principal during the telephone follow-up, with the promise that a completed Pre-Test Information Form would be forthcoming.

2.4 Participation

At the conclusion of the invitational phase, Educational Testing Service provided the U.S. Office of Education with three complete name, address, and phone number listings:

(a) A listing of all participating District Coordinators

(b) A report of the selected School Principals in each state

(c) A listing of participating School Coordinators

Copies of all pertinent listings were also provided to District Superintendents and State Coordinators.

This Final Report contains three complete numeric summaries of invitational responses. These summaries were prepared separately (by

stated for school responses in the Equating and Norming phases and for district responses. They include: the number of schools selected in the original Westat sample, the number of these schools found to be no longer in existence, the number of schools added to the Study as a separate sampling of new schools, primary sample schools unable to accept invitations to participate, and substitute schools added. The district summary contains separate data (by state) for districts with Equating schools only, with Norming schools only, and for districts with both Equating and Norming schools. District approval was received from 92.9% of the adjusted total of districts invited. Eighty-nine percent (89%) of the schools in the original sample agreed to participate. Of the total schools agreeing to participate, 99.7% actually administered tests in the Anchor Test Study and returned answer sheets and ancillary documents to Eduational Testing Service.

3. Materials Preparation and Shipping

School participation in the Anchor Test Study was confirmed by the return of an agreement postcard and a completed Pre-Test Information Form. The postcard designated a School Coordinator, the contact for all subsequent Study transactions. The Pre-Test Information Form served three primary purposes: it updated the school description file on the USOE computer tapes, it supplied additional input data required for the Study, and it provided the information needed to assign tests and prepare assignments. More specifically, the Pre-Test Information Form contained the following information:

(a) A verification of the school address, and a confirmation of the School Coordinator's name and telephone number.

(b) A selected day for testing. Participating schools were required to contact Educational Testing Service if testing was not possible during the specified administration week (April 17-21, 1972).

(c) An indication that one or more of the seven Anchor Test Study instruments was already a part of the school's testing program. This assured that Equating participants would not be assigned tests that duplicated in-school administrations.

(d) A listing of all classes at grades 4, 5, and 6. Each class was identified by a teacher name, and information was provided concerning the existence of split sessions, the number of pupils, physically handicapped or mentally retarded students, and a categorization of ability grouping.

(e) Answers to three "school description" questions, including an indication of the degree of urbanization of the community served

by the school, the general income level of the school-family population, and the socio-economic level of the school area.

Pre-Test Information Forms were received and recorded in a master control system, assuring complete returns from all participants. The forms for Equating schools were verified to assure the existence of two classes each at grades 4, 5, and 6 and/or to assure that the grades represented were the appropriate component of a combined-schools testing unit (i.e., "pseudo-school"). All Pre Test Information Forms were thoroughly edited, and School Coordinators were contacted for any missing or incomplete information.

When the editing of the Norming school Pre-Test Information Forms was completed, the forms were batched and entered into the processing system. Processing of Equating forms underwent one additional step: the designation of the two tests to be administered. According to procedures outlined by Westat Research, Inc., all equating schools were sorted into levels (matrices) for test-pair assignment. Any duplication, between the tests assigned and tests already administered as part of the school's own program, was resolved by re-assignment. These test assignments, together with the class descriptions, provided the basic information necessary for the computer preparation of control forms and shipping orders.

3.1 Packing, Shipping, and Control

For each participating school, a series of data processing forms for packing, shipping, and control were computer-produced. The Shipping Notice served as the master control form. Produced for each individual school, it listed the School Coordinator and Test Administrator kits (materials) to be packaged and shipped. The School Coordinator was

provided with a kit for each separate test form and level being administered at the school including:

(a) A School Coordinator's Manual

(b) A School Coordinator's Report Form

(c) A 5% overage of test books and answer sheets

(d) Copies of the Test Administrator's Manual

(e) Business reply envelopes for the return of answer sheets

(f) Notification postcards (acknowledging receipt of the shipment and the completion of testing)

The Test Administrator's kit provided the materials necessary to administer one test to a class or group of students. It included:

(a) A Test Administrator's Manual

(b) A Class Identification Form

(c) Test books and answer sheets

The Shipping Notice served several functions. It was used by Educational Testing Service to package and prepare the materials for the School Coordinator and Test Administrator kits. For each Coordinator/Administrator entry listed on the Shipping Notice, the specified materials were collected and packaged as a unit. The Shipping Notice assured that the correct number of kits (for the appropriate publisher's tests) were prepared for each school. One copy of the Shipping Notice was inserted in a see-through enclosure and attached to the shipping carton. This served as the mailing address for the shipment, as well as providing a receiving check-list for the School Coordinator. The Coordinators were asked to count and verify that the shipment included all the materials necessary for the test administration.

In order to identify the various individual kits within a shipment, a computer-produced label was printed to match the entries on the Shipping Notice. A label was attached to each kit, identifying it as containing a specified test for a particular teacher. In the Norming phase of the Study, each teacher (class or group test administrator) received one individually addressed kit. It included the proper quantity of materials necessary to administer the Metropolitan Reading Test to his/her class. Because two tests were administered in the Equating phase of the Study, each teacher received two personally addressed kits. While both kits were addressed to the same teacher, the labels indicated that different tests were enclosed.

Since alternate classes, within an Equating school, were asked to administer tests in an alternate sequence, it was necessary to clearly identify the test to be administered first and the test to be administered second. Therefore, a second label was affixed to each kit. It indicated whether that kit of test materials was the first or second test to be administered to the class. In order to verify that the correct labels had been applied to each kit, the kits were carefully cross-checked against the sequence of entries on the Shipping Notice.

In addition to the Shipping Notice, each School Coordinator received a computer-produced document entitled "School Coordinator's Report Form". This form was designed to provide additional testing information and to help the coordinator organize the return of completed answer sheets. It listed all the class/teacher titles and tests, individually, and suggested that the Coordinator use the form to verify complete returns. Norming School Coordinators were asked to indicate the grade level for each participating class group, the number of students in the class actually

tested, and the total number of pupils enrolled in the class. For participants in the Equating phase of the Study, the information was requested twice: once for each of the two tests administered. Equating phase participants were also asked to indicate the actual sequence of their test administrations, as well as the one test for which they prefered to have scores reported. The back of the School Coordinator's Report Form was used to record any over-all school irregularities during the test administration.

While the School Coordinator's Report Form served as a cover document for the entire school, a computer-produced "Class Identification Form" served as the cover document for each separate class group. (These forms, as well as the Pre-Test Information Form, reply postcards, and portions of the manuals and answer sheets, had been pretested by ETS and reviewed and approved by both the USOE and the Office of Management and Budget (OMB).) This Class Identification Form was included with the materials in each Test Administrator's Kit. It was completed by the Test Administrator, relayed to the School Coordinator, and returned with the completed answer sheets. The Class Identification Form's primary function was the identification of a group of answer sheets as being those of a particular class. It also served as a reminder to the Test Administrator to follow all instructions. In the Norming phase of the Study, the Test Administrator completed the Class Identification Form by recording the number of accompanying answer sheets and then returning both the answer sheets and the Form to the School Coordinator. Equating phase Test Administrators also recorded the sequence of each administration, i.e., whether the test was administered first or second. For both the Equating and Norming phases of the Study, the back of the Class

Identification Form served as a record of class or individual testing irregularities.

Each School Coordinator kit contained a 5% overage of test books and answer sheets for a particular test form and level. An extra copy of the Test Administrator's Manual was included in the kit, as well as a special manual for the School Coordinator. Instructions for the actual administration of the test, as well as Test Administrator procedures for recording information about each participating pupil, may be found in Volume XXIX of this Final Report.

Two return reply postcards were included in each school shipment. The School Coordinator was instructed to mail the first postcard to ETS upon receipt of the shipment. It served as a notification that the test materials had been delivered and that the school was supplied with all the materials necessary for testing. A control log was established at Educational Testing Service and all schools that had not forwarded their "Acknowledgement of Receipt" postcard, within approximately three weeks of the shipping date, were contacted by telephone. Most follow-up calls indicated that the shipments had been delivered, but that the cartons had not been opened (the card therefore not returned). Duplicate materials were prepared and shipped when there was the possibility of a shipment lost in transit. The School Coordinator mailed the second postcard at the same time the completed answer sheets and cover documents were released. This postcard permitted the establishment of a follow-up system for return shipments that might have been lost or delayed in the mail.

4. Material Receipt/Scoring and Reporting

The week of April 17th was designated as the primary week for schools to administer the Anchor Test Study instruments. The weeks of April 10th and April 24th were established as alternate administration weeks. Approximately 85% of the participants were able to administer the tests during the primary testing week. The alternate weeks were used by those schools with program conflicts, and several schools in the Eastern area whose semester vacations overlapped the primary administration week.

A master listing of all participating schools was maintained at ETS and, as completed answer sheets and cover documents were returned, the listing was updated to show the current status of each individual school. This listing indicated the receipt of the "answer sheets mailed" postcard, the partial return of test materials, the complete return of answer sheets and cover documents, and the scoring/reporting unit (cycle number) in which the school would be processed. This system provided an instant status report for each participating school and identified those schools with problems and/or delays. Through immediate telephone follow-up, it was possible to obtain completed answer sheets and cover documents from 99.7% of those schools that had agreed to participate.

During the course of all test material receiving, weekly status reports were prepared and released to District Coordinators, State Coordinators, and the USOE. The District Coordinator's report showed the name of each school in the District, its identification number, the name of the School Coordinator, and whether or not completed answer sheets and cover documents had been received. The State Coordinator's report summarized the number of schools with completed test materials received and the total number of schools still outstanding. State reports were combined

to present the USOE with an overall report summarizing the status of all schools.

4.1 Receiving

The initial step in material-return receiving was a verification that basic cover and identification documents had been returned. Incoming packages were scanned to ascertain that each school had enclosed the School Coordinator's Report Form (SCRF) and a Class Identification Form (CIF) for each group of answer sheets. When these documents had been identified, the completed "answer sheet groups" were matched with group identification entries listed on the SCRF and any incomplete schools were set aside to await additional shipments. Irregularity reports were scanned to determine situations that might require special handling.

When it was established that answer sheets had been enclosed for each participating class (one set of answer sheets for each class in the Norming phase and two sets of answer sheets for each class in the Equating phase) and that no irregularities existed, the answer sheets were forwarded for further scanning and editing. This more detailed editing involved a verification of the completeness and accuracy of the school's input data. The School Coordinator's Report Form, the Class Identification Form, and the answer sheets were carefully reviewed and any overall discrepancies were resolved according to the rules specified in the "Anchor Test Editing Procedures". More detailed discrepancies in counts were verified and adjusted as specified in "Anchor Test Study Receiving Rules". Individual answer sheets were numbered and a machine readable "Header Sheet" was coded for class/group identification. The "Anchor Receiving Log" was prepared, as a master control, by logging the date of answer sheet receipt,

the School Coordinator Report Form number, the school's identification number, the identification number of each class in the school, and the number of answer sheets received for each class.

4.2 Scoring and Reporting

Optical transcription of header sheet and answer sheet information was the initial data processing step in the scoring and reporting cycle. Information was transcribed directly to magnetic tape by an NCS optical transcription machine. Computer editing programs scanned the transcribed data and produced a listing of all student and/or school information that appeared to be in conflict. This information was corrected, updated, and re-introduced into the system by re-processing the original input documents. All editing specifications outlined in the "Anchor Test Editing Procedures" were implemented in order to assure "clean" and valid data. During the course of all processing, stringent quality control procedures were maintained in order to assure accuracy. These procedures are outlined in the report entitled "Quality Control Procedures for The Anchor Test Study".

While the data for all students and schools were being processed and matched to the master computer file, for use in the subsequent Norming and Equating analyses, basic score reports were provided for immediate release to the participants. Based on each publisher's current normative data, a roster of results was prepared for each school. Schools participating in the Norming phase of the Study received rosters of individual student scores for the Word Analysis, Reading, and Total score portions of the Metropolitan Reading Test. Since each student in the Equating phase of the Study had taken two tests, the school was allowed to select

the one test for which it preferred to have scores reported.

Class score rosters provided raw scores, the publisher's national percentile rank, and a stanine score for each pupil tested. For each class, the number and percentage of pupils falling within each quartile of the publisher's national norms was shown. The report also provided the class "N" (number tested), the raw score mean, and the raw score standard deviation.

In addition to the class data for each school, a combined grade level summary was prepared. This grade level report showed the raw score mean, the standard deviation, the pupil "N", and the number and percent of students falling in each quartile of the publisher's national norms (for all classes in that grade combined). In order to help schools understand the scores they were receiving, all reports included a description entitled "Interpreting Score Reports". This document served to define the scores and normative information, as well as explaining the use of summary data.

Roster preparation and release was scheduled to provide results to all schools as quickly as possible, hopefully prior to the end of the school year. More than 90% of the school reports were relased to participants during the first week in June. The reports for tardy respondents and schools with irregularities were mailed during subsequent weeks. By June 19th, all school reports had been released and all School Coordinators had been sent their honoraria for assisting with the Anchor Test Study. Upon completion of score reporting, the data analyses for the Norming and Equating phases was begun.

CHAPTER 3

SAMPLE DESIGN FOR THE ANCHOR TEST STUDY

1. Introduction

The Anchor Test project consisted of two studies: the Restandardization Study and the Equating Study. The purpose of the Restandardization Study was to establish new norms based on a national probability sample of pupils in grades, 4, 5, and 6. In the Equating Study, a separate national probability sample of pupils in grades 4, 5, and 6 was selected to establish equating relationships among the seven tests included in the study.*

The design and selection of the two samples required were done by Westat, Inc. under a subcontract with the Educational Testing Service. The design work was carried out with the cooperation of the staffs of the Educational Testing Service (ETS) and the U.S. Office of Education (OE).

This description of the sample designs for the two studies consists mostly of a statement of the procedures used, and includes only a limited discussion of why the procedures were used. This is because many of the considerations that led to the decisions involved in the determination of the particular designs used are rather complex and would require a

---

* These seven tests are: (1) the Metropolitan Achievement Test (MAT), (2) the California Achievement Test (CAT), (3) the Iowa Test of Basic Skills (ITBS), (4) the Stanford Achievement Tests (SAT), (5) the SRA Achievement Series (SRA), (6) the Sequential Tests of Educational Progress, Series II (STEP II), and (7) the Comprehensive Tests of Basic Skills (CTBS).

review of sampling theory, of relevant empirical studies, and of the basis for judgments that can only be partially documented. Such complex considerations are not discussed thoroughly in this report. However, many of the reasons that particular design decisions were made are given in this report, including relevant evidence from supporting studies. Certain general considerations are given in this section, followed by a brief summary in Section 2 of the sample designs used. In Section 3, some of the specific decisions, analyses, and considerations that influenced the choice of the sample designs are given, especially those in the more important areas where decisions were made that do not appear to follow usual sampling practice. In the remainder of this report, a detailed description of the sample design is given, including descriptions of the stratification, selection, substitution, weighting, estimation, and variance-estimation procedures.

The general goal of the sample designs for the Anchor Test study, as well as for other sample designs, was to select samples and corresponding methods of estimation such that the survey estimates would have minimum sampling errors for a given cost. An attempt was made to identify the relevant resources that were available for use in achieving an effective design within the time schedule for the study. This included a determination of the data tapes and other sources available for constructing the school sampling frame and for use in stratification and sample selection, and included identification of the relevant statistical theory and empirical knowledge acquired from past sample surveys that were available to guide

the effective application of that theory. The application of the theory involved a joint analysis of the sources of sampling variability and the sources of costs in order to arrive at an approximately optimum design.

As background for discussing the design choices that were made, it is useful to identify three types of considerations that are important in making choices to arrive at good sample survey designs. These three considerations are listed below, and some further discussion of them is given in the remainder of this section. They are:

(1) the implications of the requirement to select a probability sample;

(2) the design of the sample in an effort to minimize the sampling errors for a given cost within the time and various other limitations on the study (given the decision to use a probability-sample approach); and

(3) the specification in the sample design of only operationally feasible methods and procedures, and the introduction of controls into the design in an effort to ensure that the survey is in fact conducted in substantial compliance with the design specifications.

The first consideration followed from the specification by OE that the survey sampling errors be controlled by the design and size of the sample and measured objectively from the sample results. This requirement limits the choice of a design to one in which a random or chance procedure is used in actual sample selection, and in which all members of the population for which estimates are to be made have a known probability of selection (greater than zero). Furthermore, judgment cannot be involved in the

selection of the particular units included in the sample. Such sampling methods are frequently referred to as probability-sampling methods. Sample designs using probability sampling can yield estimates that can be termed unbiased (or unbiased in the limit) in the sense that the average value of the estimates over repeated samples (or with a large enough sample) will be equal to the value that would be obtained by applying the same measurement procedures to the entire population.

In practice, there may be minor deviations in the sampling procedures used from those of probability sampling, even though the goal is to completely achieve a probability sample. There are many alternative designs utilizing probability sampling. The other two design considerations discussed below indicate many of the factors that lead to the choice of a particular probability sample design.

The second consideration is concerned with the effort to design the sample in a way that is best for the particular study. The design decisions made ordinarily involve the use of supplemental information and always involve judgments. These judgments include decisions such as the choices of criteria to be used for stratification in the selection of the sample, the stratification priorities used and decisions made, the choices of sampling units, the use of multi-stage or single-stage sampling in obtaining a sample of schools as well as in sampling within schools, the use and choice of varying probabilities in selection, the choice of certain aspects of estimation procedures, and the use of various other design characteristics. If these design decisions are good and make effective use

of the available theory and resources, they have the effect of increasing the efficiency of the sample (i.e., of more nearly achieving minimum sampling errors for a given cost). But whether or not they are good decisions, the sample will be in no way biased by these decisions. However, proper application of relevant sampling theory (sometimes extending the theory where needed in specific instances), guided by the evidence from appropriate empirical studies and good judgment in applying them, can achieve a considerable amount or sometimes even spectacular gains in increased precision as compared with less desirable alternatives.

The third kind of consideration referred to above involves the limitation of methods and procedures specified for the survey to only those procedures that can in fact be reasonably executed as specified within the framework of the resources available for conducting the survey. Also, procedures need to be introduced to help ensure that survey performance conforms substantially to the specifications. Such considerations influence or sharply limit many of the choices of design, as indicated in some of the description that follows.

Generally, the steps taken to achieve compliance with design specifications appear to have been reasonably successful. In a practical situation, there are necessarily some defects, including some nonresponse, and perhaps a small fraction of units in the universe with no chance of being selected. However, a considerable effort was made in this study to minimize the effect of these discrepancies.

2. Brief Summary of the Sample Designs

2.1 The Sample Design for the Restandardization Study

The Restandardization sample included (1) the selection of the primary sample of 940 schools from all those listed on the public and private school universe tapes as containing at least one of the grades 4, 5, and 6, (2) the selection of a set of up to five potential substitutes for each primary selection, and (3) the selection of a supplementary sample of eligible schools omitted from the school tapes for one reason or another.

The first step in the selection of the primary sample was the assignment of a measure of size, based on the recorded enrollment in grades 4, 5, and 6, to each school in the sampling frame. Next, based on several stratification variables listed below in Section 4, the schools in the frame were grouped into 470 final strata, the total of the measures of size of the schools in each final stratum being approximately equal. Then the primary sample of 940 schools was obtained by selecting two schools from each of the 470 strata, with probabilities proportional to their measures of size. The selection of two schools per stratum was made to provide a maximum degree of stratification and still allow straightforward variance estimation. The selection of the primary Restandardization sample is described in detail in Sections 5.1 and 5.4.

The first variable that was used to stratify the schools was type of control (i.e., public or private). The public schools were further stratified by the following variables:

(1) district or county size (i.e., schools in the 50 largest school districts, other schools in large counties, and the remaining schools);

(2) percent minority of pupils or, if not available, percent minority of the community where located (eight categories);

(3) income level of the community where located (nine levels);

(4) Census region (four locations); and

(5) degree of urbanization (seven levels).

The order of priority used for these variables to define strata is approximately that indicated by the order in which they are listed above.

The private schools, which accounted for only about 11 percent of the total enrollment in grades 4, 5, and 6 of all schools in the school universe, were separated into Catholic schools and non-Catholic schools. The Catholic schools were stratified by the following variables:

(1) median income of the pupils' parents,

(2) Census region (four locations), and

(3) degree of urbanization (four levels).

The non-Catholic private schools were stratified only by geographic region.

Precise definitions and sources of the stratification variables are given later in Section 4. A detailed description of the formation of strata for the Restandardization Study is given in Sections 5.1, 5.2, and 5.3.

At the same time that 940 schools were selected for the primary Restandardization sample, potential substitutes were selected for each primary school. The first-priority potential substitute was taken to be a school in the same district with similar enrollment size and the same grade structure. The second- through the fifth-priority potential substitutes for each primary school were taken to be schools from other districts selected at random (usually from the same final stratum) with probabilities proportional to their measures of size. Substitution procedures for the Restandardization Study are described in detail in Section 5.5.

Of the 940 schools (in 703 districts) selected for the primary Restandardization sample, 826 cooperated in the study. Of the remaining 117 primary schools, 27 were ineligible for the study and therefore no attempt was made to obtain substitutes for them. For the other 87 noncooperating schools, cooperation was obtained from 80 substitutes.

In addition to the selection of the primary sample plus potential substitutes, a supplementary sample of eligible public schools omitted from the public school data tape was selected. The procedure was to first list all the public school districts that contained at least one school selected for the primary sample. Then a sample of 178 of these districts was selected

with probabilities approximately proportional to the district enrollments in grades 4, 5, and 6. Each of the selected districts was surveyed to identify those schools that contained at least one of grades 4, 5, and 6 but were omitted from the public school tape for one reason or another. All "missed" schools identified by this surveying procedure were given an appropriate chance of being added to the total sample. This supplementation procedure led to the addition of 20 missed schools to the sample, of which 12 participated in the study. There was no sample-supplementation procedure carried out for private schools. A detailed discussion of sample supplementation is given in Section 7.

All the pupils contained in the schools participating in the Restandardization Study were included in the sample. That is, there was no subsampling of pupils within schools for inclusion in the sample.

Each pupil who provided usable test data was assigned a weight to be used for estimation. Weights were based on the selection probabilities and were adjusted for school and pupil nonresponse. The determination of these weights is described in detail in Section 9.

2.2 The Sample Design for the Equating Study

The Equating sample consisted of (1) the selection of the primary sample of 448 schools (or "pseudo" schools, as described below) from those listed on the data tapes but not selected for the primary Restandardization sample, (2) the selection of a set of up to five potential substitutes for

each primary selection, and (3) the administration of a randomly assigned pair of tests (or test forms) to all, or nearly all, of the pupils in grades 4, 5, 6 in each school.

The equating procedure required that each school in the study contain at least two classes in each of grades 4, 5, and 6. Hence, before the school sample was selected, each school included in the Restandardization sampling frame and not selected for the primary Restandardization sample was checked to see whether or not the data tape indicated that it contained at least two classes in each of the grades 4, 5, and 6. If not, the school was combined with one or more schools to form a "pseudo" school, which presumably met the size requirement. Each pseudo school was treated in the sampling, selection, testing, and estimation procedures as if it were an individual school. In defining pseudo schools, an attempt was made to combine schools located near each other (i.e., in the same district if possible, otherwise in the same county). The resulting set of schools and pseudo schools comprised the sampling frame for the Equating Study. Each school or pseudo school in the frame is referred to as a "school-unit." The formation of pseudo schools is discussed more thoroughly in Section 6.1.

The equating design required 16 groups of 28 school-units for a total of 448 school-units. The 28 school-units in each group were assigned at random to the 28 possible testing assignments. These 28 assignments, illustrated in Section 8.1, consist of the 21 possible pairs of the seven tests used in the study and the seven possible assignments of two forms of a single test. The determination of the 16 groups and the test assignments of 28 school-units in each group are described in detail in Section 8.2.

The 448 school-units for the primary sample were selected from the modified frame discussed above in a way similar to the selection of the primary Restandardization sample. First, measures of size were assigned to all the school-units based on the total enrollment in grades 4, 5, and 6, as before. Then the school-units were partitioned into 224 strata, and two schools were selected from each stratum with probabilities proportional to their measures of size.

The stratification variables used were the same as those used for the Restandardization Study. For a pseudo school, the value of a stratification variable is that determined by the largest component school. The 224 final strata were formed in substantially the same way that final strata were formed for the Restandardization sample, except that the final strata were not defined to have approximately equal aggregate measures of size. This was due to the requirement that school-units falling in approximately the top sixth or bottom sixth of the socio-economic distribution be sampled at about twice the rate as the other school-units.* To achieve this double sampling rate, final strata were defined among school-units approximately in the upper and lower sixths of the socio-economic distribution to have about

---

* As discussed later in Section 6.2, the socio-economic classification of a school-unit was based on the income and percent minority measures associated with the component school with the largest enrollment in grades 4, 5, and 6. The variable sampling rate was introduced in an attempt to increase the number of relatively high and relatively low scores and consequently increase the precision of the estimated equating relationships at relatively low and high percentiles.

half the size (in terms of aggregate measures of size) as those defined in the other strata. Aside from the use of two average final-stratum sizes (rather than only one), the same procedure was used to define strata and to select two school-units per stratum for the primary sample as was used in the Restandardization Study. The definition of strata for the Equating Study is discussed further in Sections 6.2 and 6.3.

The 448 school-units selected for the primary Equating sample included 787 individual schools. Of the 787 schools, 639 cooperated in the study. With the use of 104 cooperating substitute schools, the number of cooperating school-units was increased to the target number of 448, including 743 individual schools. However, three of these 448 school-units did not provide usable data, leaving 445 school-units used in the study.

The selection of up to five school-units as potential substitutes for each primary selection was made in the same way that it was for the Restandardization Study. However, the substitution procedures used were not as straightforward as they were for the Restandardization Study. The complexity of the substitution procedures, described in Section 6.4, arose because of the requirement of two classes in each of grades 4, 5, and 6, and the use of pseudo schools.

After the primary sample was selected and the school-units were allocated to the 28 possible testing assignments, the classes in each school-unit were split into two groups, as described in Section 8.3. One group was designated to take the two tests (or test forms) in one sequence, and

the other group was designated to take them in the opposite sequence. Once these assignments were made, an attempt was made to administer the specified pair of tests in the designated sequence to all pupils in each class contained in a participating school-unit. Equating relationships were then derived from the usable test data. Also, the standard errors of the equated scores were estimated using the method of balanced half-sample replication, as described in Section 11.2.

3. Some of the Decisions, Analyses, and Considerations That Influenced the Sample Design

3.1 The Universe of Schools and the Sources Used to Construct the Sampling Frame

The target population (or universe) consisted of all pupils in grades 4, 5, and 6 in all schools, public and private, located in the 50 states and the District of Columbia, except those pupils who could not reasonably take the tests because they were handicapped or did not know English. The corresponding school universe consisted of all schools containing eligible pupils.

The sampling frame used to select the samples of schools was composed of schools listed on four school tapes. For public schools, the frame for most states consisted of those schools included on the OE 1970-1971 School Universe Tape (i.e., the ELSEGIS III C tape) which indicated the inclusion of at least one of the grades 4, 5, and 6. In some states, however,

the 1970-1971 list of schools was merged with the list for that state included on the 1969-1970 School Universe Tape. This was necessary because some of the returns to the 1970-1971 School Universe Survey had not yet been received by OE at the time that the sampling frame was constructed.*

The sampling frame for the private schools was obtained from two tapes. The first was the 1970-1971 Catholic school tape provided by the National Catholic Educational Association (NCEA).** The other tape, which listed all non-Catholic private schools in the United States in 1970-1971, was provided by OE. All schools listed on these two tapes that indicated they contained at least one of the grades 4, 5, and 6 were included in the sampling frame.

3.2 The Use of Schools as Sampling Units

In the selection of the primary samples for both studies, schools were used as the sampling units, although some consideration was given to the use of school districts in the selection procedures for public schools.

---

* A match of the 1969-1970 and 1970-1971 school tapes was performed to decide which states required the files to be merged. The states for which the merged 1969-1970 and 1970-1971 files were used for the sampling frame are Arizona, California, Connecticut, Minnesota, and South Dakota. Also, only the 1969-1970 lists were available and were used for Maine and New Jersey.

** This tape could have been obtained from OE, except that certain information needed for stratifying schools was omitted from the OE copy.

The basic alternate procedure considered, which utilized school districts as primary sampling units (PSUs), is the following. First, a stratified sample of public school districts would be selected with probabilities approximately proportional to district enrollments in grades 4, 5, and 6. The eligible schools in each of the selected districts would then be stratified by average school achievement test scores, if available. A stratified sample of schools would then be selected from each district with probabilities proportional to the total enrollments in grades 4, 5, and 6.

For such a procedure, there would have been two basic components to the total variance between school test means: the variance between district test means within a stratum of districts, and the variance between the school test means within each district. If the between-school (within districts) variance component is relatively large, this procedure could have been an effective sample design, if it were feasible to obtain the mean test scores of schools within selected districts and use this information for stratification. Related studies indicated that such mean test scores would provide highly effective stratification. However, it was not known whether or not the variance between school districts might be large enough to offset the gains from the relatively effective stratification used in the selection of schools within districts. This was one of the reasons why districts were not used as PSUs.

Perhaps an even more critical reason districts were not used as PSUs is that it was not known to what extent the selected school districts would be willing or able to provide the average school test scores for the

prior year. Finally, the time available for completion of the study did not allow for the intermediate step of selecting school districts and obtaining their cooperation in participating in the study and providing school test scores before the selection of schools for the study.

Actually, using school test scores for the selection of the school sample within the 50 largest school districts (in terms of enrollment in grades 4, 5, and 6) was seriously being considered until late in the study. That is, after the entire sample was selected without the use of school test scores, ETS requested from each of the 50 largest school districts the test means for each of the schools in the district. It was hoped that these test scores could be used to reselect the school sample from these districts. Since each of these districts had a certain or nearly certain chance to be selected, the between-district contribution to the variance would have been very minor.

Therefore, the reselection of the school sample for these districts would likely have increased sampling efficiency due to the stratification of schools within these districts. However, even though a substantial proportion of the 50 largest districts did provide school mean test scores, they were not received soon enough to redesign the sample for these districts. Also, there would have been considerable complexities involved in such a redesign at that point in the study. However, this test information is available and should be considered for possible use in future studies.

3.3 The Use of Mail and Telephone Communication With Schools and School Districts

Communication with each selected school and corresponding school district was carried out primarily by mail. That is, the cooperation of the school districts that contained schools selected for the sample was initially requested by mail. In cases for which it was necessary, mail requests were supplemented by telephone calls to the district, rather than personal visits. A similar procedure was used in contacting the schools.

A significant result of the use of mail and telephone communication was that there was no need to cluster the sampled schools in an effort to facilitate travel and reduce costs. Therefore, the schools were stratified for selection without first clustering them into geographic sampling units.

3.4 The Inclusion in the Study of All Eligible Pupils in Each Participating School

It was decided to include in the study all the eligible pupils in each school participating in either the Restandardization Study or the Equating Study. It was recognized that, from a sample-design point of view, a considerably more efficient sample (i.e., one that would yield smaller sampling errors for a given cost) could theoretically be drawn by subsampling pupils within the participating schools.

However, if the subsampling approach had been adopted, it would have exposed the process of subsampling to selection bias, which would have been difficult to control with the adopted survey approach of using local staff to select the sample and administer the tests. This could have invalidated the entire survey results. Consequently, the Educational Testing Service recommended that the system of taking all eligibles within a school be adopted, at the price of less efficient sampling but the much greater assurance of avoiding complications and biases in sample-selection procedures. They based this recommendation on their extensive experience in this area. The Westat staff concurred in the recommendation, and it was approved by the Office of Education. This decision not to subsample pupils affected the optimum allocation analysis summarized in Section 3.5 below, and described in detail in Appendix A.

3.5 The Approximate Optimum Allocation of the Restandardization Sample to Enrollment Size Classes

Because of the decision (discussed in the previous section) to include in the study all eligible pupils contained in a selected school, it is obvious that the cost for a large school in the sample would be greater than that for a smaller school due to the larger number of eligible pupils tested. Also, the larger schools would contribute more to the variance of an estimated percentile because the number of students with test scores below any specified test score would, in general, be greater for the larger schools. This indicates that a larger sampling fraction should be used for

the larger schools. Consequently, consideration was given to varying the sampling fractions by size of school in an effort to achieve an approximately optimum allocation of the Restandardization sample.

As a first step to achieve this, public schools were grouped into size classes and the optimum allocations for the size classes were computed. This analysis was based on estimating the percentile rank, P, corresponding to a particular test score, T, for a given subtest and grade. (This is shown in Appendix A to be effectively equivalent to basing the optimization analysis on the estimation of a percentile, T_P.)

As developed in detail in Appendix A, the optimum allocation for a particular size class is made by taking the sampling fraction for the size class proportional to

$$S_{Ph}/\sqrt{C_h},$$

where

C_h = the average cost of testing per school within size class h (taking account of only those costs that vary with the number of students being tested), and

S_{Ph} = the standard deviation of the Z_{hi} values among all schools in size class h.

The term Z_{hi} is the difference between the number of pupils, X_{hi}, in the i^{th} school in stratum h who have test scores below a particular score, T, and the estimated number of pupils in the school who have scores less than T, based on a regression line through the origin relating X_{hi} to the number of eligible pupils, Y_{hi}. That is,

$$Z_{hi} = X_{hi} - P\,Y_{hi},$$

where

X_{hi} = the number of eligible pupils in the i^{th} school in stratum h who have (expected) scores less than T,

Y_{hi} = the number of eligible pupils in the i^{th} school in stratum h, and

$P = \sum_{h,i} X_{hi} / \sum_{h,i} Y_{hi}$ = the percentile rank corresponding to score T.

In Appendix A, approximate values are given for the C_h values. Also, an expression is developed for the S_{Ph} term that is used to obtain approximate optimum sampling fractions for the size classes.

For a multivariate problem such as this one, the optimum allocation of the sample may, in general, be quite different for the various estimates. In this case, the various estimates are percentile ranks for various test scores (or, equivalently, various percentiles). As indicated in detail in Appendix A with the aid of some reasonable assumptions that simplify the analysis, the optimum allocation of the sample to size strata is approximately the same for the various percentile ranks (or percentiles) estimated.

Furthermore, for an optimum allocation of this type, the variances are not very sensitive to moderate variations from the optimal sampling fractions. That is, if the allocation used is moderately close to the optimum allocation, the main advantages of variance reduction are achieved. Consequently, it is reasonable to assume that the approximate optimum allocation derived for the Restandardization Study, which was based on some reasonable assumptions, was quite effective across the range of most or all of the percentile ranks and percentiles estimated.

This approximate optimum allocation analysis led to the approximate optimum sampling fractions for the Restandardization Study given in Table 3-1 in Section 5.1. As is illustrated, these sampling fractions vary from .0037 for schools with less than 50 pupils to .0356 for schools containing 500 pupils or more in grades 4, 5, and 6.

The precision of an estimated percentile rank for a given total cost was estimated for both optimum allocation and proportional allocation of the school sample to enrollment size classes. Because of the greater cost of testing in larger schools, the variance for a given total cost using optimum allocation was estimated to be only about 5 percent smaller than that estimated for the same total cost using proportional allocation. Although it was worthwhile to take advantage of this relatively small gain, there were other factors that influenced the decision to adopt the optimum allocation since it specified that higher fractions of larger schools than of smaller schools be sampled. For example, there may be some interest in analyzing the results separately for minority groups, or for schools in

large SMSAs, or for schools in central cities of SMSAs. For these and other similar purposes, the oversampling of large schools relative to small schools with a larger total number of students tested seemed to be more appropriate. These considerations, in addition to the modest gains in sampling efficiency, were the basis for the recommendation to adopt the optimum allocation of the school sample to enrollment size classes.

3.6 The Use of Measures of Size to Allocate the School Sample to Enrollment Size Classes

The approximate optimum allocation analysis carried out in Appendix A determined the sampling fractions used for the enrollment size classes in the selection of the Restandardization sample. Originally, it had been planned to use the enrollment size classes to define strata in order to apply these approximately optimum sampling fractions.

However, a number of recent studies indicate that there is substantial variation between schools in the performance of students on achievement tests.* The studies indicate that the variation between school test means accounts for as much as one-third of the total variance among individual

---

* The following two references indicate that the total variance among pupil test scores accounted for by schools is about 25 to 35 percent:

(1) George Mayeske, et al., A Study of Our Nation's Schools (a U.S. Office of Education working paper analyzing the data collected for the Coleman report).

(2) Robert L. Thorndike, ed., Educational Measurement (2nd edition), American Council on Education (1971), Chapter 15.

test scores. Therefore, effective stratification of schools can substantially reduce the variance of sample estimates of percentile ranks or percentiles. Characteristics such as the average educational or income level of parents of students and the fraction of students in the school that belong to minority groups are among the most effective in accounting for the variation between school means.* These characteristics are shown to be more important than school size in accounting for the variation between school means.

Consequently, it was decided not to stratify schools by size class, but to control the allocation to size classes, to a large extent, by assigning approximate measures of size to schools by enrollment class, and by selecting schools with probabilities proportional to these measures of size. This allowed more extensive use of other variables to stratify schools for sample selection. Therefore, this sampling approach provided an allocation to size classes nearly equal to the approximate optimum allocation derived, and still provided for further gains resulting from the extensive use of income and percent minority measures in the stratification of schools. In Section 4, all the variables used for stratification are defined and the sources of these variables are given.

---

* The references cited on the previous page and the following two references show high correlations between average school test scores and socio-economic variables relating to the surrounding community and the families of the students:

(1) Thomas P. Hogan, Socioeconomic Community Variables as Predictors of Test Performance, Ph.D. dissertation, Fordham University (1969).

(2) Richard M. Jaeger, Designing School Testing Programs for Institutional Approval: An Application of Sampling Theory, Ph.D. dissertation, Stanford University (1970).

The measures of size assigned to the various enrollment classes were based on the approximate optimum sampling fractions determined in Appendix A. The derivation of the set of measures of size from these sampling fractions is given in Appendix B. It should be noted that the simplifying assumptions made in the optimization analysis that determined the measures of size to be assigned to the schools had no effect on the unbiased character of the sample estimates. That is, although these assumptions led to the probabilities of selection, unbiased estimates can be made from a sample for any arbitrarily assigned (nonzero) probabilities of selection. However, the sampling errors of the estimates can be substantially affected by the probabilities of selection. That is, if the probabilities of selection are assigned in an approximately optimum manner, as was accomplished in this case, sampling errors can be lower than those that would be obtained using an alternate set of selection probabilities.*

It was decided to use the same measures of size to select the Equating sample as were used to select the Restandardization sample. This decision was based on the analysis presented in Appendix A which indicated that the optimum allocation of the school sample to enrollment size classes for the Equating Study should be the same (adjusted for the total sample size desired) as for the Restandardization Study.

---

* The optimum allocation analysis was based on variance and cost estimates made in advance of the study. The quality of these estimates and the resulting sample allocation can, and should, be examined through analyses of the actual survey data. Such an evaluation would be useful for future designs of this type.

3.7 The Inclusion in the Sample of All of the Pupils Contained in Grades 4, 5, and 6 in Each Participating School

In both the Restandardization Study and the Equating Study, all the eligible grades (4, 5, and 6) were included in the sample within each participating school. The primary goal was to provide norms and equated scores for each grade separately. For this purpose, the clustering of the grades within a single school as a sampling unit did not increase the variance of estimated percentile ranks or percentiles for a particular grade, even though the scores for the grades were positively correlated. On the contrary, a positive correlation in average test scores between grades within a school has the effect of reducing the variance of estimated differences between grades in the standardized scores.

The positive correlation in test scores between grades within a school arises primarily because the different grades serve the same communities, and because the same students continue or flow from one grade to the next within the same school. However, in many school systems, grades 4, 5, and 6 are not all contained in the same school. For example, several small schools containing grades 4 and 5 could be supplying sixth-grade students for one or two large schools.

Because of these types of arrangements, some consideration was given to selecting the Restandardization school sample to include approximately the same flow of grades in separated schools as is automatically achieved in schools that contain the three eligible grades. The purpose of this would

have been to further reduce the sampling errors of the differences in standardized test scores for the successive grade levels. This approach was rejected mainly due to the design complexities that would have resulted. Also, the necessary information on the grade flows between schools not containing all of the three eligible grades was not readily available. Furthermore, for the sample size being obtained, the sampling errors were quite small and any additional gains in reduced variances of differences that could have been achieved by this approach were small at best.

3.8 The Possible Use of School District Information from the 1970 Census to Improve Sample Design

As a result of special arrangements made by the Office of Education, the 1970 Census is providing, for the first time, information on characteristics of the population within school districts. This information was under preparation at the time of selection of the samples for the Anchor Test Study. It was therefore thought that this information might be available in time to be used in the estimation procedures as a means of introducing post-stratification. However, it could not be used because availability was delayed.

The principal relevance of this information is its availability for future analysis and use to the extent that it can increase the effectiveness of future sample designs. Its potential for such use should be investigated.

4. The Stratification Variables Used in the Study

4.1 Introduction

In general, the stratification variables and corresponding procedures used to define strata should be chosen to achieve as much homogeneity as feasible of schools within strata with respect to the characteristics being estimated, and thus minimize the contribution of the variability within strata to estimates of the proportion of students who score below various test scores. Also, depending on the number of strata to be defined, the number of variables that can be used for stratification is limited. These considerations led to the choices of the stratification variables listed previously in Section 2.1. The priority of these variables used to define strata was also indicated in that subsection. If the choices of variables and corresponding priorities for stratification are not as good as they could have been, the variances of sample estimates will be somewhat larger than they otherwise would have been. However, as indicated previously, no bias would be introduced as a result of such choices.

In the following subsections, the actual measures and sources used for the stratification variables are given. A detailed description of how these variables were used to define strata is given in Sections 5.1, 5.2, 5.3, 6.2, and 6.3.

4.2 Type of School Control (Public or Private)

The schools were grouped for stratification into public schools and private schools, depending on the school tapes on which they were listed. The public schools accounted for about 89 percent of the total enrollment in grades 4, 5, and 6. The Catholic schools contained just under 10 percent of the enrollment, and the non-Catholic private schools contained just over 1 percent of the enrollment.

4.3 Size of School District and Population of County

The public schools were divided into groups for stratification, indicated by the following breakdown:

(1) schools contained in the 50 largest districts in terms of enrollment in all grades; and

(2) other schools--

  (a) schools in counties with a 1960 Census population of 50,000 or more, and

  (b) schools in counties with a 1960 Census population of less than 50,000.

The names of the 50 largest districts were obtained from an OE list containing all school districts ordered by total enrollment. The 1960 Census population for counties used to divide the schools in the other districts into those in large and small counties was obtained from a county tape compiled by Westat from Census data.

Although the breakdown by county size of the schools located in districts other than the 50 largest probably added a worthwhile contribution to the stratification of schools, this breakdown was made primarily because of the two different income measures used to stratify schools. That is, as indicated in the next subsection, income measures associated with schools in large counties were different from those used in small counties. Furthermore, these two measures were not comparable. Therefore, strata were defined separately among schools in large and small counties.

4.4 Income Class

It was difficult to find an appropriate income measure to associate with a particular school. Data that gave the reported average income of the students' parents for a particular school were available and were used for the Catholic schools, but were not available for the public schools or the other private schools. Furthermore, it was not feasible to obtain income measures that corresponded closely to school attendance areas. After some limited investigation, two income sources were used to assign measures to public schools. Measures obtained from these sources take either the ZIP Code area or the county containing the school as an approximation to the school attendance area, as described below.

For schools contained in counties with a 1960 Census population of 50,000 or more, it was decided to base the school income measure on 1966 tax returns included on an IRS tape. That is, using this tape, the measure

assigned to such a county was the average adjusted gross income based on tax returns with the same five-digit ZIP Code as that for the school.*

In the investigation of the available income measures, it was discovered that, for schools located in counties with 1960 Census income of less than $50,000, the ZIP Code incomes were somewhat erratic and often not very indicative of the economic status of the school attendance areas. Therefore, for schools located in these counties, the income measure associated with a school was taken to be the 1960 Census median income of the county containing the school.

For incomes based on the ZIP Code areas (i.e., for schools in large counties), the following eleven income classes in terms of average adjusted gross income were used:

---

* In order to avoid giving a few extremely large incomes too much influence on this value, the average adjusted gross income for a ZIP Code area was defined as

$$(X_1 + N_2\bar{X}_2)/N ,$$

where

X_1 = the aggregate of the adjusted gross incomes under $10,000 for the five-digit ZIP Code area,

N_2 = the income of taxpayers reporting adjusted gross income of $10,000 or more,

$\bar{X}_2$ = the average adjusted gross income of these N_2 taxpayers, or $15,000, whichever is less, and

N = the total number of taxpayers in the five-digit ZIP Code area.

(1) Less than $2,000
(2) $2,000 - $2,999
(3) $3,000 - $3,999
(4) $4,000 - $4,999
(5) $5,000 - $5,999
(6) $6,000 - $6,999
(7) $7,000 - $7,999
(8) $8,000 - $8,999
(9) $9,000 - $9,999
(10) $10,000 and over
(11) Not available

The following eleven classes in terms of median family income were used for income measures obtained from the 1960 Census (i.e., for schools located in small communities):

(1) Less than $1,000
(2) $1,000 - $1,999
(3) $2,000 - $2,999
(4) $3,000 - $3,999
(5) $4,000 - $4,999
(6) $5,000 - $5,999
(7) $6,000 - $6,999
(8) $7,000 - $7,999
(9) $8,000 - $8,999
(10) $9,000 - $9,999
(11) Not available

For Catholic schools, an estimated average income of the parents of the students was included on the Catholic school tape provided by NCEA.

Based on these incomes, the following eleven classes were used in the stratification procedures for Catholic schools:

(1) Less than $3,000
(2) $3,000 - $4,999
(3) $5,000 - $6,999
(4) $7,000 - $8,999
(5) $9,000 - $9,999
(6) $10,000 - $11,999
(7) $12,000 - $13,999
(8) $14,000 - $15,999
(9) $16,000 - $19,999
(10) $20,000 and over
(11) Not available

For the non-Catholic private schools, an appropriate income measure was not available.

4.5 Percent Minority

There were no appropriate percent minority measures available for private schools. For public schools, the following eight levels of percent minority were used:

(1) Under 5 percent
(2) 5 - 9 percent
(3) 10 - 19 percent
(4) 20 - 39 percent
(5) 40 - 59 percent
(6) 60 - 79 percent
(7) 80 percent and over
(8) Not available

Measures of percent minority were obtained for many of the schools from a tape provided by the Office of Civil Rights. For schools that had filed with the Office of Civil Rights and were included on the tape, the percent minority was computed as the percent of students reported as Negro, Spanish-surnamed American, or American Indian. For schools that had not filed with the Office of Civil Rights, the percent minority was computed as the percent nonwhite for the corresponding county from the Westat county tape. The percent nonwhite recorded on this tape for counties was obtained from the 1970 Census.

4.6 Geographic Region

In the sample-stratification procedures, the four Census geographic regions were used for both public and private schools. The region associated with a school was determined by the state code given on the corresponding school tape. The four regions and an indication of the states included in each of these regions are the following:

(1) Northeast--Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, and Pennsylvania;

(2) North Central--Ohio, Indiana, Illinois, Michigan, Wisconsin, Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebrasks, and Kansas;

(3) South--Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida, Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiana, Oklahoma, and Texas; and

(4) West--Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada, Washington, Oregon, California, Alaska, and Hawaii.

4.7 Degree of Urbanization

In the stratification procedures, a degree of urbanization code was used for public schools and Catholic schools. No such code was used for the non-Catholic private schools.

After excluding the 50 largest school districts, the following seven levels of degree of urbanization were used for the remaining public schools:

(1) schools in the (next) 100 largest school districts that were also contained in standard metropolitan statistical areas (SMSAs), and

(2) schools in other SMSAs

The remaining schools were given a degree of urbanization code based on the percent urban of the county containing the school. The levels of percent urban used for these counties are the following:

(3) 75 percent or more urban
(4) 50 - 74 percent urban
(5) 25 - 49 percent urban
(6) Less than 25 percent urban
(7) Not available

The 51st through 150th largest school districts were identified by the same OE list that was used to identify the 50 largest districts. The percent urban for each county was taken from the Westat county tape. These percent urban figures were obtained for this tape from the 1970 Census.

For Catholic schools, four levels of degree of urbanization were used in the stratification procedures. These four levels, which were indicated on the Catholic school tape provided by NCEA, are the following:

(1) schools in urban areas but not in an inner city,
(2) schools in an inner city,
(3) schools in suburban areas, and
(4) schools in small towns or rural areas.

5. Stratification and Selection Procedures for the Restandardization Study

5.1 Basic Stratification and Selection Design

Based on the approximate optimum allocation analysis discussed in Section 3.5 above and described in detail in Appendix A, a measure of size was assigned to each school. That is, measures of size were assigned to the schools in such a way that the expected sampling fractions for the enrollment classes would be approximately equal to those determined by the approximate optimum allocation analysis. The table below gives these measures of size along with approximate population sizes, sample sizes, and sampling fractions for the schools in each enrollment category. The calculation of the assigned measures of size from the approximate optimum allocation is described in Appendix B.

Table 3-1. Measures of Size

Enrollment in Grades 4, 5, and 6	Approximate Number of Schools in Universe	Approximate Sample Size	Approximate Sampling Fraction	Assigned Measure of Size
Less than 50	11,537	43	.0037	4
50 - 99	10,584	93	.0088	9
100 - 199	18,667	270	.0145	14
200 - 499	22,347	487	.0218	22
500 or more	1,322	47	.0355	35
Total	64,457	940		

* The first two columns in this table were obtained from the 1968-1969 School Universe Tape. In cases where the enrollment was not given grade by grade, the enrollment in grades 4, 5, and 6 was estimated as:

$$(g/t)e \ ,$$

where

g = the number of grades 4, 5, and 6 in the school,

t = the number of grades 1 through 12 in the school, and

e = the enrollment in the grades 1 through 12 in the school.

After these measures of size were assigned to the schools in the universe, the schools were grouped into 470 (final) strata in two stages. First, the schools were partitioned into 47 "major" strata using the variables (1) type of control (public, Catholic, non-Catholic private), (2) size of school district or county, (3) income level of the school or school attendance area, and (4) percent minority of the pupils or the school attendance area.

Then, each of the 47 major strata were subdivided into final strata using the variables (1) Census geographic region, and (2) degree of urbanization of the surrounding area. On the average, ten final strata were defined per major stratum, totaling 470 final strata. Also, the final strata were defined so that the totals of the measures of size of the schools in the final strata were approximately equal from final stratum to final stratum. This two-stage stratification process is described in more detail in the next two subsections.

From each of the 470 final strata, two schools were selected randomly with probability proportional to their measure of size (PPS). This provided the primary Restandardization sample of 940 schools. At the same time the primary sample was selected, several potential substitutes were picked with PPS for each of the schools selected for the primary sample. The selection of the primary sample plus substitutes is described later in detail in Sections 5.4 and 5.5.

5.2 The Definition of the 47 Major Strata

A number that was used in defining the major strata was the overall average of the aggregate measure of size, $\bar{A}$, of a final stratum. Letting A represent the aggregate measures of size of all schools in the entire sampling frame, $\bar{A}$ was derived as follows:

$$\bar{A} = A/470 = 1,041,289/470 = 2,215.5 .$$

Major strata were defined separately for public and private schools. For public schools, the following steps were used to define major strata. First, the public schools were placed in three groups:

(1) schools in the 50 largest school districts based on district enrollment in all grades,

(2) those schools remaining that are in counties with a 1960 population of 50,000 or more, and

(3) all other schools (i.e., schools in counties with a 1960 population of less than 50,000).

Then, within each of these three public school groups, the schools were placed in the cells of an 8 x 11 table, based on the eight percent-minority categories described in Section 4.5 and the eleven income classes described in Section 4.4. There was a total of 264 cells (3 x 8 x 11) in the three tables, many of them containing no schools. The sum of the measures of size of the schools in each cell was then computed. Next, the 47 major strata were formed from these 264 cells by combining adjacent cells in the same table, where necessary, to provide groups of cells with aggregate measures of size averaging about 10 $\bar{A}$. These major strata were defined so that, to an extent feasible, they contained schools with similar income and percent-minority classifications. Also, an effort was made to define major strata having aggregate measures of size that were close to whole integer multiples of $\bar{A}$. This was done so that the final strata defined in different major strata would have approximately the same aggregate measures of size. Though the aggregate sizes of the major strata varied considerably, they were defined to have a minimum aggregate size of about 5 $\bar{A}$, except in one case for which the aggregate size was approximately 4 $\bar{A}$. The number of major strata defined for public schools was 40.

For private schools, major strata were defined separately for Catholic schools and for the remaining private schools. There were only enough non-Catholic private schools to form one major stratum. The Catholic schools were grouped into six major strata using the estimated average income reported on the NCEA tape of the parents of the pupils in each school. This was done by first placing schools in the eleven income classes indicated in Section 4.4 for Catholic schools. Then, major strata were defined by combining adjacent income categories, using criteria similar to those used to collapse the 264 public school cells into 40 major strata. That is, schools were grouped in such a way that each major stratum contained schools with similar income measures. Also, major strata were defined to have a minimum aggregate measure of size of about 5 $\bar{A}$, except in one case for which the aggregate measure of size was about 4 $\bar{A}$.

The aggregate measures of size for all of the 47 major strata are given in Table 3-2 in the next section. In Appendix C, the aggregate measures of size are given for each of the 264 cells for the public schools and eleven cells for Catholic schools, along with an indication of how the cells were collapsed into the 47 major strata.

5.3 The Definition of Final Strata

The first step in defining final strata within each of the 47 major strata was to determine the number of final strata to be defined in each major stratum. If A_M represents the aggregate measure of size for major

stratum M, then the number of final strata, S_M, defined in major stratum M was computed as follows:

$$S_M = A_M/\bar{A} = A_M/2,215.5 \text{ , rounded to the nearest integer.} \qquad (3\text{-}1)$$

As described above, an attempt was made to define major strata in such a way that $A_M/\bar{A}$ would be close to an integer. The values of S_M for all the major strata are given in Table 3-2 below.

Next, the average size, $\bar{A}_M$, of a final stratum defined in major stratum M was computed. The average was computed as simply the total major stratum size divided by the number of final strata to be defined:

$$\bar{A}_M = A_M/S_M \, . \qquad (3\text{-}2)$$

The values of $\bar{A}_M$ varied somewhat from the overall average, $\bar{A}$, depending upon the rounding involved in computing S_M by equation 3-1. The values of $\bar{A}_M$ are given in Table 3-2.

Table 3-2. Major and Final Stratum Sizes for the Restandardization Study

Major Stratum Number (M)	Aggregate Measure of Size (A_M)	Number of Final Strata (S_M)	Average Final Stratum Size ($\bar{A}_M$)
1	19,339	9	2,149
2	17,982	8	2,248
3	17,905	8	2,238
4	10,262	5	2,052
5	14,437	7	2,062
6	11,567	5	2,313
7	13,583	6	2,624
8	28,756	13	2,212
9	14,773	7	2,110
10	12,352	6	2,059
11	49,575	22	2,253
12	9,395	4	2,349
13	31,089	14	2,221
14	11,579	5	2,316
15	12,311	6	2,052
16	12,860	6	2,143
17	11,541	5	2,308
18	11,476	5	2,295
19	13,268	6	2,211
20	10,531	5	2,106
21	20,144	9	2,238
22	17,473	8	2,184
23	14,856	7	2,122
24	48,050	22	2,184
25	13,201	6	2,200
26	27,678	12	2,307
27	34,594	16	2,162
28	41,174	19	2,167
29	142,365	64	2,224
30	17,021	8	2,128

Table 3-2 (Continued)

Major Stratum Number (M)	Aggregate Measure of Size (A_M)	Number of Final Strata (S_M)	Average Final Stratum Size ($\bar{A}_M$)
31	12,669	6	2,112
32	54,257	24	2,261
33	20,372	9	2,264
34	15,043	7	2,149
35	15,520	7	2,217
36	18,232	8	2,279
37	15,277	7	2,182
38	22,815	10	2,282
39	12,791	6	2,132
40	11,157	5	2,231
41*	21,273	9	2,127
42**	17,361	8	2,170
43**	47,191	21	2,247
44**	13,137	6	2,190
45**	11,123	5	2,225
46**	12,112	5	2,422
47**	1,822	4	2,456
Total	1,041,289	470	

* Stratum consisting of non-Catholic private schools.

** Strata consisting of Catholic schools.

Once the number of final strata, S_M, and the average size, $\bar{A}_M$, of a final stratum were obtained for major stratum M, the final strata were then defined in such a way that they had approximately equal aggregate measures of size. This was done by listing the schools in the major stratum in order by geographic region and within geographic region by the degree of

urbanization code defined in Section 4.7 above.* The following ordering scheme was used:

(1) schools in the Northeast region ordered from most urban to least urban,

(2) schools in the North Central region ordered from least urban to most urban,

(3) schools in the South region ordered from most urban to least urban, and

(4) schools in the West region ordered from least urban to most urban.

Opposite ordering by urbanization code (i.e., alternating between most urban to least urban and least to most) from one region to the next was used to provide continuity of the degree of urbanization ordering from region to region. Next, the cumulative measure of size, cum (A_{Mi}), was computed for each school on the major stratum list. Then, each school on the list that had a cumulative measure of size less than or equal to $\overline{A}_M$ was assigned to the first final stratum defined in major stratum M. The next school on the list was also included in the first final stratum if its measure of size, A_{Mi}, and a cumulative measure of size, cum (A_{Mi}), satisfied the following inequality:

$$\text{cum } (A_{Mi}) - \overline{A}_M \leq A_{Mi}/2 \ .$$

---

* Since there were no urbanization codes available for non-Catholic private schools, the schools in the major stratum containing non-Catholic private schools were listed within geographic region by the order in which they appeared on the data tape.

Of the remaining schools in major stratum M, all those that had a cumulative measure of size less than or equal to $2\ \bar{A}_M$ were assigned to the second final stratum defined in major stratum M. In addition, the next school on the list was included in the second final stratum if its measure of size and cumulative measure of size satisfied the following inequality:

$$\text{cum } (A_{Mi}) - 2\ \bar{A}_M \leq A_{Mi}/2 \ .$$

The other final strata defined in major stratum M were formed in a similar way. The resulting set of final strata defined in a particular major stratum had nearly equal aggregate measures of size.

5.4 Selection of the Primary Sample for the Restandardization Study

After the 470 final strata were defined, two schools were selected from each stratum with probability proportional to size (PPS). The resulting sample of 940 schools constituted the primary Restandardization sample.

The selection of the primary sample was carried out in each major stratum, say, major stratum M, as follows. First, the schools in each of the final strata in major stratum M were placed in random order, based on a computer-generated random number. Next, new cumulative measures of size, cum (A_{Mi}), were computed for all the schools in the major stratum M, based on their new position on the list resulting from the randomization of schools within final strata. These cum (A_{Mi}) values were computed without regard to final-stratum boundaries.

The sample was then selected systematically from the major stratum as follows. First, the selection interval used was $\bar{A}_M/2$. Then, the initial selection number was taken as $\bar{A}_M/4$, rather than being picked at random. This initial selection number was approximately one-fourth of the aggregate measure of size of the first final stratum in major stratum M. All the other selection numbers for major stratum M were obtained from this initial number in the standard way--sequentially adding the selection interval, $\bar{A}_M/2$, to the previous selection number until the aggregate measure of size for the major stratum had been reached.

This procedure produced the following set of $2\ S_M$ selection numbers:

$$\bar{A}_M/4,\ \ 3\ \bar{A}_M/4,\ \ 5\ \bar{A}_M/4,\ \ \ldots,\ \ (4\ S_M - 3)\bar{A}_M/4,$$
$$(4\ S_M - 1)\bar{A}_M/4\ .$$

These $2\ S_M$ selection numbers identified the $2\ S_M$ schools for the sample from major stratum M (with two schools from each of the S_M final strata) in the usual way. That is, the school determined by a particular selection number, k, was the first school on the major stratum list whose cumulative measure of size was greater than or equal to k. The two schools selected from a particular final stratum were listed about 1/4 and 3/4 of the way down the final-stratum list, in terms of cumulative measures of size. This selection scheme is illustrated in Figure 3-1 in the next subsection.

This particular selection procedure was used for several reasons. First, the selection numbers were generated systematically rather than randomly to provide an easy way to have PPS selection and sampling without replacement. Second, even though the aggregate measures of size varied slightly from final stratum to final stratum, the same selection interval was used throughout the major stratum, rather than allowing it to vary slightly from one major stratum to the next, to provide a straightforward selection and weighting procedure. The use of a varying selection interval would have complicated the weighting procedures since the selection would have been exactly PPS only within a final stratum and approximately PPS throughout the major stratum. (Since the selection interval was fixed at $\overline{A}_M/2$, the selection probability for the i^{th} school in major stratum M was $2\ A_{Mi}/\overline{A}_M$, regardless of the size of its final stratum.)

The use of a fixed initial selection number rather than a random one eliminated the possibility of selecting only one school from some final strata and three from other final strata. (As mentioned earlier, the selection of <u>two</u> schools per final stratum allowed straightforward variance estimation and a high degree of stratification.) The use of a fixed start did not bias the selection in any way, since the schools within each final stratum were ordered randomly prior to the selection of the sample.

In brief, the selection procedure used provided an easy way to select exactly two schools with PPS and without replacement from each final stratum.

5.5 The Selection and Use of Substitutes for the Primary Restandardization Sample

Of the primary sample of 940 schools, there were 27 schools selected for the sample that were not part of the school universe. In most cases these were schools that were closed or that did not contain any of the grades 4, 5, and 6. Since these schools were not part of the universe, they were simply eliminated from the sample without obtaining substitutes for them. (Also, no adjustments in the weighting procedures were made for the elimination of the 27 schools from the sample.)

An intense effort was made to secure the cooperation of the 913 schools remaining in the primary sample. Because of this effort, all but 87 of the 913 schools agreed to cooperate in the study. As imputation for these 87 nonrespondents, an attempt was made to obtain substitutes for each of them.

Potential substitutes for all the schools in the primary sample were selected at the same time the primary sample was selected. The first-priority substitute for a school in the primary sample was a school, if any, that (1) was located in the same school district; (2) contained the same composition of grades 4, 5, and 6; (3) belonged to the same major stratum; and (4) had the same or an adjoining measure of size as the school in the primary sample. In cases for which more than one school fulfilled these criteria for a particular school in the primary sample, the one that had a recorded total enrollment in grades 4, 5, and 6 closest to that of the primary school was taken as the first-priority substitute. No other school

in the same district as the primary school was allowed to be a possible substitute for that school.

There were up to four other potential substitutes selected for each primary school. These were selected with PPS, mostly from the same final stratum as the primary school. The selection numbers used to identify these substitutes were obtained as follows. If R_i represents the selection number for the i^{th} school selected in the primary sample from major stratum M, the following six selection numbers were used to obtain a maximum of four potential substitutes (the second through the fifth priorities) for the i^{th} selected school:

$$R_i - .1\,\bar{A}_M\,,\quad R_i + .1\,\bar{A}_M\,,\quad R_i - .2\,\bar{A}_M\,,\quad R_i + .2\,\bar{A}_M\,,$$

$$R_i - .3\,\bar{A}_M\,,\quad R_i + .3\,\bar{A}_M\,.$$

That is, the second-, third-, fourth-, and fifth-priority substitutes for the i^{th} school selected were the first four schools that were eligible of the six schools determined by the six selection numbers given above. Only those schools not selected for the primary Equating sample and those contained in the same district as the i^{th} selected school were ineligible to be used as these lower-priority substitutes.

This procedure for selecting potential substitutes is depicted in Figure 3-1 below. The asterisks labeled A and B represent the two selection numbers used to select the primary sample from the first final stratum in major stratum M. The points labeled A1 through A5 represent the five selection numbers used to obtain the second-, third-, fourth-, and fifth-priority

substitutes for the primary school determined by selection number A. (There is no point A6 since substitutions across major stratum boundaries were not permitted.) Similarly, points B1 through B6 determine the second- through fifth-priority substitutes for the school determined by selection number B. Points C, D, C1 through C6, and D1 through D6 have similar representations for the second final stratum in major stratum M. In some cases, a school was a low-priority potential substitute for two primary schools. This occurred because of the overlap of substitute selection numbers illustrated in Figure 3-1.

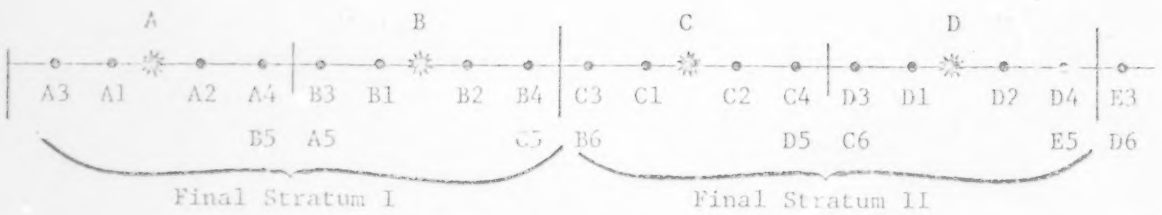

Figure 3-1. Representation of the scheme used to determine the selection numbers for the primary and substitute selections

The effort to obtain substitutes for the 87 noncooperating primary schools resulted in the cooperation of 80 substitute schools. In nearly all cases, cooperation was obtained from either the first- or second-priority substitutes.

6. Stratification and Selection Procedures for the Equating Study

6.1 The Definition of Pseudo Schools

The selection of the sample of schools for the Equating Study was similar in many respects to that for the Restandardization Study. One substantial difference was caused by the decision that schools in approximately the upper and lower sixths of the socio-economic distribution be sampled at twice the rate that the other schools were sampled. The resulting design modifications are discussed in the next section.

Another difference in the selection procedure for the Equating Study occurred because of the requirement that each school in the sample contain at least two classes in each of grades 4, 5, and 6. It was not feasible to satisfy this requirement by selecting the sample in a way similar to the Restandardization sample, and then supplementing the sample in some way to provide extra classes for those selected schools lacking the minimum number of classes. The reason that this was not feasible is that it would have been impossible to compute accurate selection probabilities. This minimum size requirement had to be met by grouping some schools together prior to the selection of the Equating sample. That is, each school in the sampling frame that indicated that it did not contain at least two classes in each of grades 4, 5, and 6 was combined with at least one other school to form a "pseudo" school, which did contain the required number of grades and classes.

The following procedure was used to form pseudo schools. First, the 940 schools selected for the primary Restandardization sample were removed from the sampling frame. Then, the public schools remaining in the frame that were incomplete (i.e., had insufficient grades or class sections) were combined with other public schools in the same district, preferably, or in the same county, if necessary, to form complete pseudo schools. Private schools in the frame that did not meet the size requirement were combined with other private schools in the same county, preferably, or in the same state, if necessary.

For public schools, an attempt was made to combine incomplete schools in the same district that complemented each other. For example, schools containing only grade 4 were combined with schools containing only grades 5 and 6. For incomplete schools for which a perfect complement could not be found, a search was made to find an incomplete school in the same district that was nearly a complement. For example, a school with only grades 5 and 6 might be combined with one containing only grades 4 and 5. It was hoped that pseudo schools formed in this way would group schools together that service the same pupils as they progress through grades 4, 5, and 6. After all possible combinations of this type were made among incomplete schools within a district, combinations of the same type were made among schools within the same county. After this, any incomplete schools

remaining were combined with other complete schools or pseudo schools in the same district or, if necessary, the same county.*

For private schools, a similar procedure was carried out for defining pseudo schools by combining incomplete schools in the same county or, when necessary, in the same state.

Pseudo schools were treated the same as individual (complete) schools in the stratification, selection, and testing procedures. Measures of size for pseudo schools were determined as they were for other schools (based on Table 3-1) using the total of the enrollments in grades 4, 5, and 6 of all the component schools of the pseudo school. All the stratification characteristics of a pseudo school, such as income level, percent minority, and degree of urbanization, were determined by the component school with the largest enrollment in grades 4, 5, and 6. Those individual schools and pseudo schools that made up the sampling frame for the Equating Study are referred to as school-units. There was a total of 40,554 school-units in the Equating sampling frame. Of these, 18,696 were pseudo schools, while 21,858 were individual schools.

6.2 The Definition of the 27 Major Strata

Once the pseudo schools were defined, a sample of 448 school-units was selected for the Equating Study--two school-units with PPS from each of

---

* In a very few instances, all the schools in a county did not form a complete pseudo school. In such a case, the pseudo school remained in the sampling frame as an incomplete pseudo school, and the deficiencies were taken care of by the substitution procedures if the pseudo school was selected.

224 final strata. As was done in the Restandardization Study, the strata were defined in two steps, using the same variables to define major and final strata as those used in the Restandardization Study. For the Equating Study, 27 major strata were defined, and an average of just over eight final strata were defined per major stratum to produce the 224 final strata required for the Equating sample.

Because of the requirement that school-units in approximately the upper and lower sixths of the socio-economic distribution* be sampled at twice the rate as the other school-units, two average final stratum sizes, $\bar{A}_1$ and $\bar{A}_2$, were needed in defining major and final strata. These were obtained as follows. The total of the measures of size of all school-units in the Equating sampling frame was 831,547. The upper and lower socio-economic sixths accounted for about one-third of this aggregate measure of size (i.e., 277,182). To double the sampling rate for schools in these two-sixths, the same number of final strata, 112, was defined among schools in these two-sixths as among schools in the remaining four-sixths. The approximate average aggregate measure of size, $\bar{A}_1$, for a final stratum defined in either the upper or lower sixth of the socio-economic distribution was therefore computed as follows:

$$\bar{A}_1 = 277,182/112 = 2,475 .$$

---

* The socio-economic classification of a school was determined by the income and percent minority categories associated with it. The use of these measures to group schools into the upper and lower sixths and middle four sixths of the socio-economic distribution is described later in this subsection.

For strata defined among the other schools, the approximate average aggregate measure of size, $\bar{A}_2$, was twice $\bar{A}_1$. That is,

$$\bar{A}_2 = 2(2,475) = 4,950 .$$

The steps used to define major strata were very similar to those used for the Restandardization Study given in Section 3.2, except for the use of the two average final stratum sizes. Major strata were defined separately for public and private school-units. Public school-units were placed in the same three groups that were used for the Restandardization Study. The school-units in each of these three groups were placed in an 8 x 11 table, based on the percent minority and income measures, as before.

Next, each of these three 8 x 11 tables was partitioned into three sets of cells. One set contained school-units in approximately the top sixth of the socio-economic distribution of the school-units in the table in terms of measure of size. For example, if one of these tables had an aggregate measure of size of 60,000, cells corresponding to the highest income classes and lowest percent-minority categories would be grouped together in such a way that their aggregate measure of size was about 10,000. Similarly, the lower socio-economic sixth would be formed by grouping cells corresponding to low-income and high percent-minority categories. The third set of cells would be the remainder of the table--those cells corresponding approximately to the middle four-sixths of the socio-economic distribution.

In the cells corresponding to the upper and lower sixths of each table, major strata were defined by grouping adjacent cells in such a way that the aggregate measures of size of the major strata would be approximately equal to multiples of $\bar{A}_1$. The major strata were defined to average about 10 $\bar{A}_1$, with a minimum major stratum size equal to approximately 5 $\bar{A}_1$. Major strata in the middle four-sixths were defined by grouping adjacent cells in such a way that the sizes of the major strata would be approximately equal to multiples of $\bar{A}_2$. The aggregate sizes of major strata in the middle four-sixths had, on the average, slightly lower aggregate sizes, varying from about 3 $\bar{A}_2$ to 10 $\bar{A}_2$. There were 22 major strata defined for public school-units.

For private schools, major strata were defined separately for Catholic school-units and non-Catholic school-units. As in the Restandardization Study, there were only enough non-Catholic private school-units to form one major stratum. This major stratum was assigned to the top socio-economic sixth. Four major strata were defined among Catholic schools, using the estimated parents' income for the school-units to divide them into the upper sixth, lower sixth, and middle four-sixths, and to define the major strata.

The aggregate measures of size for all of the 27 major strata are given in Table 3.3 in the next section. In Appendix C, the aggregate measures of size are given for each of the cells in each of the tables for both public and private schools, along with an indication of how the cells were collapsed into major strata.

6.3 The Definition of Final Strata and Selection of the Primary Sample

The final strata were defined within major strata for the Equating Study in almost exactly the same way that they were for the Restandardization Study. This procedure was described in Section 5.3 above and will not be repeated here. The only variation of the procedure described in Section 5.3 was in the use of equation 3-1, which was used to determine the number of final strata defined in a given major stratum. The term $\bar{A}$ in equation 3-1 was replaced by $\bar{A}_1$ or $\bar{A}_2$, depending on whether the major stratum was in the upper or lower sixth or the middle four-sixths of the socio-economic distribution.

Table 3-3 below gives the number of final strata defined in each major stratum and the average aggregate measure of size of the final strata defined in each major stratum.

As indicated earlier, pseudo schools were treated the same way that single schools were treated in the selection of the sample. The selection of the primary sample of 448 school-units for the Equating Study was carried out in each major stratum in exactly the same way as it was in the Restandardization Study. This was described earlier in Section 5.4 and will not be repeated here. Counting all the component schools separately, the sample of 448 schools or pseudo schools included a total of 787 schools for the primary Equating sample.

Table 3-3. Major and Final Stratum Sizes for the Equating Study

Major Stratum Number (M)		Aggregate Measure of Size (A_M)	Number of Final Strata (S_M)	Average Final Stratum Size ($\bar{A}_M$)
1	Upper Sixth of Socio-Economic Distribution	19,363	8	2,420
2		32,010	13	2,462
3		29,655	12	2,470
4*		14,314	5	2,863
5		40,418	16	2,526
6**		16,676	7	2,382
7		17,802	3	5,934
8		18,019	4	4,505
9		24,023	5	4,805
10*		22,993	5	4,599
11		19,565	4	4,891
12		51,519	10	5,152
13		39,632	8	4,954
14		65,65[illegible]	13	5,050
15		49,902	10	4,990
16		48,940	10	4,894
17		34,692	7	4,956
18		24,928	5	4,986
19		15,876	3	5,292
20		29,734	6	4,956
21		45,665	9	5,074
22*		36,276	7	5,182
23	Lower Sixth of Socio-Economic Distribution	29,932	12	2,494
24		19,335	8	2,417
25		32,018	13	2,463
26		39,890	16	2,493
27*		12,740	5	
Total		831,547	224	

* Strata consisting of Catholic schools.

** Stratum consisting of non-Catholic private schools.

6.4 The Selection and Use of Substitutes for the Primary Equating Sample

Five potential substitutes were selected for each school-unit in the primary Equating sample. These potential substitutes were selected in exactly the same way that potential substitutes were selected in the Restandardization Study. This was described in Section 5.5 above and will not be repeated here. Pseudo schools were treated the same as single schools in constructing these lists of potential substitutes. The district associated with a pseudo school was taken to be the district corresponding to the component school with the largest enrollment in grades 4, 5, and 6. (In most cases, all the components of a pseudo school were located in the same school district.)

Although the formation of lists of potential substitutes was straightforward, the actual substitution procedures were somewhat complex for the Equating Study due to the presence of pseudo schools. First, in some instances selected school-units that supposedly contained at least two classes in each grade did not actually satisfy this size requirement. Such a school was still retained in the sample, but had to be supplemented by a school from the list of substitutes. The first school on the list of substitutes that contained the missing classes and agreed to participate in the study was used to supplement the deficient primary school. In cases in which the substitute school-unit was a pseudo school, only those components necessary to make the deficient primary school complete were added to the sample.

In some instances, the school-unit selected for the primary sample was a pseudo school and only a portion of the component schools declined. In a number of such cases, the remaining (i.e., cooperating) component school or schools still contained the required number of classes. In that event, no substitution was used to replace the component school or schools that declined. On the other hand, in cases in which the remaining component schools did not constitute a complete school-unit, the same substitution procedures were used as those described above for deficient primary schools.

The substitution procedures described above for dealing with the school-units selected for the primary sample were also used for dealing with substitute school-units.

Since the test equating procedure required 448 school-units with two classes in each of grades 4, 5, and 6, an attempt was made to obtain substitutes for all of the school-units in the primary sample that did not cooperate in the study. This included primary schools that were closed or did not contain any of the grades 4, 5, or 6. These types of losses plus the refusals totaled 155 schools, counting the component schools individually. In an attempt to bring the total number of cooperating school-units to 448, 104 individual substitute schools were added to the study. As a result, all but three of the 448 school-units provided usable test data.

7. Selection of a Supplemental Sample of Public Schools Omitted From the Sampling Frame for the Restandardization Study

7.1 Summary of the Sample Supplementation Procedure

After the primary Restandardization and Equating samples were selected, a procedure was carried out to supplement the primary Restandardization sample with a sample of the relatively small number of public schools omitted from the sampling frame for one reason or another. Originally, it was hoped that both the Restandardization and Equating samples would be supplemented. However, it was subsequently decided that it would not be feasible to supplement the Equating sample. The main reason for this decision was the requirement for the Equating Study that the sample contain exactly 448 schools, two from each of 224 strata. Therefore, if a supplementary Equating sample had been selected, the schools chosen would have served only as possible substitutes. Also, the requirement of two classes in each of grades 4, 5, and 6 would have complicated the selection of a supplementary Equating sample. Still another complication would have been the need to assign supplementary schools to either the top sixth, bottom sixth, or middle four-sixths of the socio-economic distribution.

The supplementary Restandardization sample was selected in two stages. A sample of 178 school districts was selected as the first-stage sample. Each district selected was asked by telephone to check the list extracted from the School Universe Tape of all schools in the district containing grades 4, 5, and 6, and identify any eligible schools omitted from

this list. The second stage of selection consisted of subsampling the lists of missed (i.e., omitted) schools. This supplementary procedure provided a sample of 20 missed schools.

The sample supplementation procedure did not include private schools. This was because there were no primary sampling units for private schools that could easily be used to identify missed schools (i.e., such as school districts). At one point during the survey, some investigation was done regarding the possibility of supplementing the sample of private schools. A procedure considered for Catholic schools was the use of Catholic dioceses in a way similar to the use of public school districts. However, since these dioceses are quite large, it may not have been feasible for them to check for missed schools. Also, a list of all dioceses in the United States was not readily available. Therefore, since there was not much time available to carry out additional sample supplementation, this procedure was not adopted. Another procedure that was considered briefly was to survey the public school districts selected for supplementation for private schools as well as public schools. However, it would have been rather difficult, especially in the time available, to obtain lists of the private schools included in certain public school districts. It would also have been a problem to check such lists against the two private school tapes to see which ones were listed. Therefore, this procedure was also rejected.

7.2 The Selection of a Sample of 172 Public School Districts for Sample Supplementation

For the selection of the supplementary sample of missed schools, the Office of Education specified that a sample of between 150 and 200 public school districts be surveyed for missed schools. It was decided to select this sample from the set of the 703 districts represented in the primary Restandardization sample. In terms of sampling error, perhaps a somewhat more efficient district sample could have been obtained by selecting the sample of districts from all districts in the frame rather than from only those represented in the primary sample. However, there was not enough time available to contact districts for the first time to seek their participation in the study. The use of a sample of districts already contacted was more expedient, and district cooperation for this supplementary procedure was probably better than it would have been otherwise. The steps that were used to select this sample of districts are described below.

First, after the selection of the Restandardization sample, a list was made of the 703 public school districts that contained at least one school selected for the primary Restandardization sample. Then, for each of these districts, the probability that at least one of its schools would be selected into the sample was estimated. This probability, which is denoted by P_d, is referred to as the initial district-selection probability. The method used to estimate the P_d values is given in Appendix D.

Before selection of the sample, the districts were grouped into strata, based on geographic region and size as indicated by the initial district-selection probability, P_d. Then, two districts (or more in four strata)* were selected at random from each of the strata. Two districts were selected per stratum, rather than one, to provide for variance estimation. Since only those districts represented in the primary sample had a chance to be selected for sample supplementation, missed schools in new districts not represented in the school universe file had no chance of selection in the supplementary sample.

The districts were stratified by geographic region and the P_d values in conjunction with the sampling fractions given below in Table 3-4. (The derivation of these sampling fractions is discussed later in this subsection.) This was done by first ordering the districts in each of the four regions by the P_d values. Then, strata were defined by grouping the appropriate number of districts together within each region that have similar values of P_d. For example, within each region, districts with values of P_d between .9 and 1 were grouped into strata containing six districts so that, with two districts selected per stratum, the sampling rate would be one in three. Of course, the number of districts in this P_d-interval was not always six or a multiple of six in each region. Therefore, some of the districts in this interval had to be grouped together with districts

---

* The reason there were more than two districts selected from four strata was that the smallest districts had to be subsampled at a 100-percent rate, as indicated later in this subsection.

Table 3-4. District Subsampling Fractions and Selection Probabilities

Initial District Selection Probability (P_d)	Corresponding Estimated Average District Enrollment (thousands)	Subsampling Fraction (f_{2h})	District-Selection Probabilities for Supplementation ($P_{d'}$)
1	50	1/2	.5
.9 - .999	35	1/3	.300 - .333
.8 - .899	25	1/4	.200 - .225
.7 - .799	18	1/6	.117 - .133
.6 - .699	13	1/7	.0833 - .1000
.5 - .599	9.5	1/7	.0714 - .0833
.4 - .499	7	1/7	.0571 - .0714
.3 - .399	5	1/7	.0429 - .0571
.25 - .299	3.9	1/6	0417 - .0500
.20 - .249	3.2	1/6	.0333 - .0417
.15 - .199	2.5	1/6	.0250 - .0333
.10 - .149	1.6	1/6	.0167 - .0250
.075 - .0999	1.2	1/5	.0150 - .0200
.05 - .0749	.9	1/4	.0125 - .0188
.04 - .0499	.7	1/4	.0100 - .0125
.03 - .0399	.6	1/3	.0100 - .0133
.02 - .0299	.5	1/3	.0067 - .0100
.01 - .0199	.4	1/2	.0050 - .0100
Less than .01	.2	1	Less than .01

in the P_d-interval of .8 to .899. In general, districts in adjacent P_d-intervals were sometimes included in the same stratum. However, only in a few instances were strata defined that contained districts from more than one region. This was done only to avoid grouping districts with P_d values of 1 with other districts, and to avoid placing districts that had substantially different sizes in the same stratum.

The above process led to the definition of 82 strata of districts. As indicated previously, two districts were chosen at random from each of these 82 strata, except in four of the strata that had subsampling rates of 1. As a result, 178 districts were selected from the 82 strata, 22 coming from the four strata that had the 100-percent subsampling rate.

The district subsampling rates given in Table 3-4, which played an important role in the definition of district strata described above, were based on two considerations. One was the specification that between 150 and 200 districts be selected. The other was an attempt to use approximately an optimum allocation of the district sample to the 19 size-class intervals listed in Table 3-4. That is, since in many cases the standard deviation on survey items among sampling units is roughly proportional to the size of the sampling units, it was decided to select districts with overall probabilities of selection for supplementation approximately proportional to their total enrollments in grades 4, 5, and 6. Also, since experience from other studies indicates that standard deviations would tend to be greater than proportional to size for the smaller districts, and to

some extent for the larger districts, the subsampling rates were set to produce overall district-selection probabilities somewhat higher than proportional to size for the larger and smaller districts.

The procedure used to determine the district subsampling rates to conform to the considerations mentioned above was rather complex and included some personal judgment. First, the overall probability that school district d would be selected for the supplementation procedure, P'_d, is equal to the product of the initial district-selection probability for the primary Restandardization sample, P_d, and the subsampling rate for the P_d interval containing district d, f_{2h}. That is,

$$P'_d = P_d f_{2h} . \qquad (3-3)$$

Since the P_d values were known approximately, the P'_d values were determined by the choices of the interval subsampling rates, f_{2h} values. In order to choose the f_{2h} values to make the P'_d values approximately proportional to district enrollments, the relationship between district enrollments and the P_d values was estimated. This was done by plotting a large number of points relating district enrollment to P_d for a sample of districts represented in the primary sample. The approximate relationship observed led to the correspondence between P_d-intervals and district enrollments given above in Table 3-4. For example, the average district enrollment in grades 4, 5, and 6 of districts with initial selection probabilities between .8 and .9 was estimated to be 25,000.

Once this approximate relationship between the P_d-intervals and district enrollments was established, the f_{2h} values given in Table 3-4 were determined (on a somewhat trial-and-error basis), which provided a sample of 178 districts and produced overall district-selection probabilities approximately proportional to district enrollment. Also, as indicated earlier, the subsampling rates for smaller and larger districts were set to provide overall district probabilities somewhat higher than proportional to their enrollments.

7.3 The Selection of the Sample of Missed Schools from the Districts Selected for Supplementation

Each of the 178 school districts selected for the sample supplement was surveyed to determine if it contained any eligible schools omitted from the School Universe Tape. To do this, lists were prepared of all schools contained in each of the 178 selected districts as indicated on the School Universe Tape. Each district was contacted by telephone and requested to verify the list of schools for that district. Except for very large districts, an attempt was made to obtain the names, addresses, and enrollments in grades 4, 5, and 6 of missed schools over the telephone. The very large districts were asked to send a list of missed schools by mail.*

---

* Due to some reorganization of school districts, some consolidation of schools, and other factors, it was sometimes difficult to determine whether a particular school was actually a missed school. For example, a missed school for a particular district may have been included on the School Universe Tape as belonging to another school district. Since it was not feasible to search the entire tape for a particular school, some judgments had to be made in identifying missed schools.

Even those districts that refused to allow their schools to participate in the study were encouraged to help survey their districts for missed schools. As a result of their effort, all districts selected for sample supplementation did participate in the district surveying process. This procedure yielded a total of 138 missed schools identified by the districts.

An attempt was made to give each missed school listed an overall selection probability approximately equal to what it would have been if the school had been included in the original sampling frame. This was done as follows. First, the measure of size, A_i, for each missed school was determined from the enrollment in the same way that it was done for the other schools (i.e., using Table 3-1). (The school enrollments were often not provided and the measures of size had to be approximated, based on the recorded enrollments of the other schools in the same district.) In order to provide an easy way to produce the appropriate overall school-selection probabilities, an adjusted measure of size, A'_i, was then computed from the original measure of size, A_i, as follows:

$$A'_i = A_i / P'_d \quad . \tag{3-4}$$

All missed schools with adjusted measures of size greater than 1100 were included in the supplementary sample with certainty. All other missed schools were listed district by district. Next, cumulative adjusted measures of size, cum (A'_i), were computed for each school on this list. A systematic sample of these missed schools was then selected by first picking

the initial selection number as a random number, R, between 1 and 1100. The selection numbers for the sample supplement were then computed as follows:

$$R,\ R + 1100,\ R + 2200,\ R + 3300,\ \ldots\ .$$

These selection numbers identified the schools selected for the sample supplement in the same way that the primary sample was selected. That is, the school determined by a particular selection number was the first missed school on the list that had a cumulative adjusted measure of size, cum (A'_i), equal to or greater than the selection number. This procedure led to the selection of 20 schools for the supplementary sample, including six sub-sampled with certainty.

The selection probability, P'_i, for a missed school contained in district d, except for missed schools selected with certainty, is given approximately by the following equation:

$$\begin{aligned} P'_i &= P(\text{district d is selected}) \\ &\quad \times P(\text{school i is selected given that district d is selected}) \qquad (3\text{-}5) \\ &= P'_d \frac{A'_i}{1100} = P'_d \frac{(A_i/P'_d)}{1100} = \frac{A_i}{1100}\ . \end{aligned}$$

Since the selection intervals used in the selection of the primary Re-standardization sample (i.e., the $\bar{A}_M/2$ values) averaged about 1100 or so, the selection probability for a missed school given above was approximately equal to what it would have been if the school had not been omitted from the School Universe Tape.

Schools subsampled with certainty have selection probabilities equal to the district selection probability. That is, for a missed school contained in district d and selected from the lists of missed schools with certainty, its selection probability is simply given as follows:

$$p'_i = P'_d . \qquad (3\text{-}6)$$

Of the 20 schools selected for sample supplementation, seven refused to participate in the study. In such cases, it was hoped that substitute missed schools could be obtained. Unfortunately, six of the seven schools were from districts that were too small to allow substitutes. That is, there were no available missed schools contained in districts that were close to the sizes of the two districts containing the six missed schools. Also, there was not enough time to obtain a substitute for the other noncooperating school. Furthermore, one of the 13 schools that agreed to participate in the study did not provide usable test data. Therefore, of the 20 missed schools selected to supplement the Restandardization sample, 12 actually participated in the study.

8. The Assignment of Tests to Schools for the Equating Study

8.1 The Possible Test Assignments

The Equating Study involved the estimation of equating relationships among seven tests. The test design required that either two separate tests or two parallel forms of the same test be assigned to each school in

the sample. Therefore, there were 28 possible test assignments for a given school: 21 involving two different tests, and 7 involving parallel forms of the same test. In each school, one of the tests assigned was administered first (usually during a morning session) and the other test (or test form) was administered second (usually during an afternoon session). Half (or nearly half) of the classes in each grade in a school took the tests in one sequence, while the other classes took the tests in the opposite sequence.

The total number of possible test assignments for a particular class in a school was therefore 56 since there were 28 possible pairs of tests or test forms and two possible testing sequences for each pair. These 56 test assignments are depicted in the 7 x 7 test assignment matrix in Table 3-5. Each cell represents the two tests identified by the corresponding row and column for the cell. The 42 cells off the main diagonal represent the 21 possible pairs of tests, while the 14 half-cells on the main diagonal correspond to test assignments involving two different forms of the same test.

In addition to identifying a pair of tests (or test forms on the diagonal), each cell (or half-cell) represents a particular testing sequence. For off-diagonal cells, the test identified by the corresponding row is designated as the test taken first, while the other test is designated as the one taken second. For a cell on the main diagonal, the half-cell above the diagonal represents one testing sequence of the two parallel forms of the particular test, and the half-cell below the diagonal represents the opposite testing sequence.

Table 3-5. The Test Assignment Matrix

Test Administered First \ Test Administered Second	CAT	CTBS	ITBS	MAT	STEP II	SRA	SAT
CAT	1	2	3	4	5	6	7
CTBS	2	8	9	10	11	12	13
ITBS	3	9	14	15	16	17	18
MAT	4	10	15	19	20	21	22
STEP II	5	11	16	20	23	24	25
SRA	6	12	17	21	24	26	27
SAT	7	13	18	22	25	27	28

The cells and half-cells in Table 3-5 are numbered from 1 to 28, indicating the 28 possible test pairs. Each number appears in two cells (or two half-cells on the main diagonal). One of the two cells (or half-cells) identifies one of the possible testing sequences for the particular pair of tests (or test forms), and the other cell (or half-cell) identifies the other possible testing sequence. For example, the cell labeled "10" that is above the main diagonal in the test-assignment matrix given in Table 3-5 corresponds to the test pair "CTBS, MAT." The test sequence determined by this cell is CTBS first and MAT second. The cell labeled "10" in the lower half of the matrix corresponds to the same test pair, but the opposite testing sequence.

8.2 The Assignment of the Schools in the Primary Equating Sample to Test Assignment Matrices

Each of the 448 school units contained in the primary Equating sample was assigned to a pair of tests (or test forms) by the allocation of schools to cells in test assignment matrices. The sample size of 448 school-units was used to provide 16 replications of assignments to the 28 pairs of cells or half-cells in a test matrix. Three basic steps were involved in making the assignments of the 448 school-units to the 16 matrices.

First, the 448 schools were partitioned into eight groups, referred to as principal strata. Each of these principal strata consisted of 56 school-units--two from each of 28 final strata. The grouping of the 224 final strata into sets of 28 to define principal strata corresponded to the selection from

the top sixth, bottom sixth, and middle four-sixths of the socio-economic distribution in the way illustrated in Table 3-6 below.

Table 3-6. The Grouping of Final Strata Into Principal Strata

Portion of Socio-Economic Distribution	Approximate Number of Final Strata	Approximate Number of Principal Strata
Top Sixth	56	2
Middle Four-Sixths	112	4
Bottom Sixth	56	2
Total	224	8

Table 3-6 shows that two groups of 28 final strata were defined among schools selected from those in approximately the top socio-economic sixth, two groups from those in approximately the bottom sixth, and four groups from those in the middle four-sixths. Actually, 61 (rather than 56) final strata were defined among schools estimated to be in the top socio-economic sixth of the socio-economic distribution. Therefore, five of the final strata considered as part of the top sixth were assigned to a principal stratum in the middle four-sixths. Also, 54 (rather than 56) final strata were defined among schools estimated to be in the lowest socio-economic sixth. Consequently, two final strata from those considered as part of the middle four-sixths were placed in one of the two principal strata defined for the bottom sixth.

In defining these principal strata, an attempt was made to order the final strata by their socio-economic classification. This was done by first ordering the major strata from highest socio-economic classification to lowest, based on the income and percent-minority measures. Then, the final strata were listed, major stratum by major stratum, from those contained in the major stratum with the highest socio-economic rating to those contained in the major stratum with the lowest rating. Within major strata, the final strata were listed simply in the order in which they were defined. Then, the groups of 28 final strata were defined straight from the order in which they were listed. That is, the first 28 final strata listed made up the first principal stratum. The next 28 listed formed principal stratum number two. In a similar way, all eight principal strata were defined.

After the principal strata were formed, the next step in assigning tests to schools was to divide each of the principal strata of 56 school-units into two sets of 28 school-units. This was done by randomly assigning one of the two school-units selected from each of the 28 final strata to one of the two sets and the other school-unit selected from the final stratum to the other set. After this was done, each principal stratum consisted of two sets of school-units, each set containing one school-unit from each of the 28 final strata included in the principal stratum.

The total number of sets of 28 school-units defined in this way was 16. The reason these 16 sets of school-units were defined in pairs (as described above) rather than individually was to allow for estimating

the variances of the equating relationships. These variances were estimated using McCarthy's balanced half-sample replication method, as described in the final section of this chapter, Section 11.*

The third step in assigning school-units to test pairs was to assign the 28 school-units in each of the 16 sets to the cells in the test assignment matrix illustrated in Table 3-5. This was done by using a random permutation of the numbers from 1 to 28 to assign a random number between 1 and 28 to each school-unit in a set. This random number became the number of the two cells or half-cells to which a school-unit was assigned.

As indicated previously, there were 104 individual schools used as substitutes for the Equating Study. The test assignment for a substitute school was taken to be that of the primary school-unit for which the school served as a substitute.

8.3 The Assignment of Classes to Test Sequences

Once a school-unit was assigned to a cell in the test matrix, half (or about half) of the classes were assigned to the top half (and the corresponding testing sequence), while the remainder of the classes were assigned to the lower half (and the opposite testing sequence). This was done as follows.

---

* For a discussion of the balanced half-sample approach to variance estimation, see Philip J. McCarthy, Replication: An Approach to the Analysis of Data from Complex Surveys, Cornell University, a publication of the U.S. Department of Health, Education, and Welfare, Public Health Service, National Center for Health Statistics, Series 2, No. 14, April 1966, pp. 13-18 especially.

First, the eligible classes in a particular grade within each school-unit were numbered consecutively beginning with "one." Whenever ability groupings were indicated on the Pretest Information Forms, classes were numbered as follows: above average first, average second, no indication third, and below average fourth. By design, there were at least two classes in each grade contained in each participating unit.

Then, the odd-numbered classes in the school-unit were assigned at random to either the cell (or half-cell) in the upper half of a test assignment matrix or the corresponding cell (or half-cell) in the lower half of the matrix. The even-numbered classes were assigned, of course, to the corresponding cell (or half-cell) in the opposite half of the matrix.

It was originally planned to assign classes to testing sequences independently from school to school within a test matrix. However, it was decided that steps should be taken to prevent a particular test from being taken by the odd-numbered classes first in more than half of the six school-units assigned to pairs involving that test. It was believed that, because of the numbering of classes by ability groupings, a test would tend to have higher scores if it were taken first (i.e., usually during the morning session) by a majority of the odd-numbered (i.e., higher ability) classes. Therefore, classes were assigned to the top and bottom half of each test matrix by selecting at random one of the two testing sequence patterns illustrated in Table 3-7 below.

Table 3-7. Testing Sequence Assignment Patterns

Tests

Tests	1	2	3	4	5	6	7
1	1 / 0	0	1	0	1	0	1
2	1	0 / 1	0	1	0	1	0
3	0	1	1 / 0	0	1	0	1
4	1	0	1	0 / 1	0	1	0
5	0	1	0	1	1 / 0	0	1
6	1	0	1	0	1	0 / 1	0
7	0	1	0	1	0	1	1 / 0

Tests

Tests	1	2	3	4	5	6	7
1	0 / 1	1	0	1	0	1	0
2	0	1 / 0	1	0	1	0	1
3	1	0	0 / 1	1	0	1	0
4	0	1	0	1 / 0	1	0	1
5	1	0	1	0	0 / 1	1	0
6	0	1	0	1	0	1 / 0	1
7	1	0	1	0	1	0	0 / 1

In these testing sequence assignment patterns, a "0" in a cell indicates that the even-numbered classes are to be assigned to that cell (or half-cell on the diagonal), while a "1" indicates that the odd-numbered classes are to be assigned to the cell. Of course, cells that correspond to the same two test combinations have opposite numbers since the classes in a school-unit assigned to the top half of the matrix must be administered tests in the order opposite from that designated for the classes assigned to the bottom half. Also, in each row and column of the matrices, there are exactly three off-diagonal 1's and three off-diagonal 0's. Therefore, for the off-diagonal cells, each test will be administered to the odd-numbered classes first in three school-units and to the even-numbered classes first in three school-units. Also, since the two test assignment patterns are exact opposites, there is a .5 probability that the even-numbered classes in a particular school-unit will be assigned to the top half of a matrix and a .5 probability that the odd-numbered classes in that school-unit will be assigned to the top half.

8.4 Resolving Conflicts of Some of the Test Assignments

It was stipulated that a school could not, as part of the Equating Study, administer a test to its pupils that had been given during the previous year. Therefore, since it was known that schools in New York were being given the Metropolitan Achievement Tests (MAT) and that schools in Delaware were being given the Sequential Tests of Educational Progress, Series II (STEP II), test assignments were made that avoided MAT for New York and STEP II for Delaware.

Conflicts of this type for schools in other states were not discovered until after the original assignments were made. All tests administered by a school during the preceding year were noted on their Pretest Information Forms. At the time that this form was received from a school, the test assignments for that school could be checked for conflicts. In cases for which no conflicts existed, schools were sent the test materials for the original test assignment.

Whenever conflicts occurred, an attempt was made to resolve them by interchanging test assignments of pairs of school-units assigned to the same matrix. The first type of interchange made was between two or more school-units that all had test conflicts that could be resolved by the interchange. Whenever this type of switch was not available, an attempt was made to interchange the test assignment for a particular school with that of another school in the same matrix that had not returned its Pretest Information Form. A random procedure was used to select a school for this purpose.

When nearly all the Pretest Information Forms had been returned, there was a rather limited opportunity to resolve conflicts. In two cases, the test assignment of a particular school had to be interchanged with that of a school assigned to a cell in the "companion" test assignment matrix. (The companion matrix was the one used for test assignments in the other half of the same principal stratum.)

9. The Assignment of Weights to the Respondents

9.1 Summary of the Weighting Procedures

In general, weights were computed as inverses of selection probabilities with adjustments for nonresponse. Each weight contains a school-weight component and a pupil-weight component.

The school weights were computed the same way for both the Restandardization and Equating Studies.* These components consist of the product of the inverse of the school-selection probability and a small adjustment for school nonresponse. In the next section the method used to compute the school weights is given.

The expression for the pupil-weight component varies slightly for the two studies. For the Restandardization Study, this expression consists of only a pupil nonresponse adjustment. There is no pupil-selection factor since all pupils in a given school were included in the sample. Pupil nonresponse adjustments were made by imputing to the nonrespondents the responses of those with characteristics similar to the nonrespondents. This was done by defining weighting classes based on the variables sex, race, and IQ, and making weight adjustments within these classes. A description of how this was done is given in Section 9.3 below.

---

* It was planned that the appropriate adjustment in the school weights for the Equating Study would be made to allow for the elimination of the schools selected for the primary Restandardization sample from the sampling frame for the Equating Study. However, this adjustment was not made, since it would have been quite small due to the small sampling fractions used to select the Restandardization sample, and since it would have added considerable complexity to the weighting procedures due to the use of pseudo schools.

For the Equating Study, the expression for the pupil-weight component includes the same pupil nonresponse adjustment as was used for the Restandardization weights. However, there are two additional weighting factors for Equating Study respondents. One factor reflects the random assignment of the classes in a school to a particular testing sequence. The other factor is an adjustment to equalize the sum of the pupil weights in each of the 56 cells (or half-cells) indicated in Table 3-5 in Section 8.1. These weight adjustments are described in Section 9.3 below.

9.2 School Weights

The same expression for school weights was used for both the Restandardization Study and the Equating Study. This expression was applied to each school-unit (i.e., school or pseudo school) that actually participated in either study. In the case of a substitute school, the school weight was computed as though the substitute had been selected originally as part of the primary sample. That is, substitute schools participating in the studies were treated the same way that participating schools contained in the primary samples were treated.

Pseudo schools were weighted as a unit in the same way that individual schools were weighted. In cases in the Equating Study involving partial substitution needed to meet the requirement of at least two classes

in each of grades 4, 5, and 6, the school weights were based on the school or pseudo school (i.e., the school-unit) that was actually used in the study. In general, for the purposes of weighting schools in both studies, all participating schools or school-units were treated as if they were selected as part of the primary sample, regardless of the substitution involved.

The school weights were computed in two steps. First, the initial or sampling weight, which is simply the inverse of the school-selection probability, was computed. The expression used to compute this initial weight, W_{Mi}, for the i^{th} school or school-unit participating from major stratum M is the following:

$$W_{Mi} = 2\,\bar{A}_M/A_{Mi}\,, \qquad (3\text{-}7)$$

where

$\bar{A}_M$ = the average size of a final stratum defined in major stratum M as before, and

A_{Mi} = the measure of size of the i^{th} school or school unit participating from major stratum M.

Each school weight was adjusted for nonresponse within the major stratum. This nonresponse adjustment is simply the sum of the school weights among all schools selected from major stratum M divided by the same sum among the participating schools from major stratum M. The resulting

expression used to compute the adjusted school weight, W'_{Mi}, for the i^{th} school participating from major stratum M is as follows:

$$W'_{Mi} = W_{Mi}\left(\frac{\sum_{i}^{T_M} W_{Mi}}{\sum_{i}^{T'_M} W_{Mi}}\right), \qquad (3\text{-}8)$$

where

$T_M = 2\,S_M$ = the number of schools selected from major stratum M, and

T'_M = the number of schools participating from major stratum M.

For schools selected in the sample supplement for the Restandardization Study, the expression used to compute the initial sampling weight, W_{Mi}, varied slightly from that given in equation 3-7. For weighting purposes, the supplementation sample was treated as a sample selected from an additional major stratum (i.e., major stratum number 48) for the Restandardization Study. The following expression was used to obtain the weighting expression, W_{48i}, for the initial sampling weight of the school selected into the sample supplement:

$$W_{48i} = \begin{cases} 1{,}100/A_{48i} \text{ if the school was not subsampled from the list of missed schools with certainty,} \\ 1/P'_d \text{ if the school was subsampled with certainty;} \end{cases} \qquad (3\text{-}9)$$

where

A_{48i} = the measure of size of the i^{th} participating school in the sample supplement, and

P'_d = the estimated probability that the school district containing the i^{th} selected school (i.e., district d) was selected to be surveyed for missed schools.

The nonresponse adjustment for the weights of schools participating in the sample supplement was computed the way that it was for the main sample. Therefore, the adjusted school weights, W'_{48i} values, were computed using equation 3-8 above, with the index "M" taken to be "48."

The nonresponse adjustments from major stratum to major stratum were very small, except for the supplementation stratum. For this stratum, the nonresponse adjustment factor was 1.89. However, for the Restandardization sample (excluding the supplementary sample of missed schools), these adjustment factors were 1.0 in all but six major strata. For these six strata, the adjustment factors ranged from 1.01 to 1.16. For the Equating Study, the adjustment factors were 1 except in two major strata, for which they were 1.03 and 1.11.

9.3 Pupil Weights for the Restandardization Study

Since all the pupils in each participating school were included in the sample, there is no pupil-selection component in the pupil weighting expression. Therefore, the weighting expression used to compute pupil

weights for the Restandardization Study consists only of the product of the school weight, W'_{Mi}, and a pupil nonresponse adjustment that accounted for absenteeism and unusable test data. The pupil nonresponse adjustments, described below, were made separately within weighting classes.

Weighting classes were defined using the following variables: race, sex, and IQ. They were defined separately for each major stratum and without regard to the school attended.* For a particular major stratum, weighting classes were formed by first setting up a three-way table based on the categories of the variables race, sex, and IQ, and placing each pupil who provided usable test results in the appropriate cell in the table.** Any cell that contained at least 20 respondents was taken as a weighting class.*** An attempt was made to collapse cells that did not meet this condition with adjacent cells in such a way that the condition would be met by the collapsed cells.

---

* Weighting classes were not defined separately for each school, since this would have severely limited the extent to which the variables race, sex, and IQ could have been used to define weighting classes. Although there is certainly some similarity among pupils, which is accounted for by the school they attend, it was believed that the variables race, sex, and IQ were considerably more important in grouping responding and nonresponding pupils together who would have similar responses.

** There were, of course, missing race, sex, and IQ items for some pupils. Therefore, the tables used to define weighting classes had to contain cells that allowed for various combinations of complete and missing items.

*** The choice of a minimum of 20 was somewhat arbitrary. It was believed that 20 respondents would be enough to provide nonresponse adjustments that were not erratic and would still permit the grouping into weighting classes of pupils with quite similar race, sex, and IQ characteristics.

The collapsing procedure was carried out in each major stratum as follows. First, cells were combined across various IQ classes. In cases in which this did not produce combined cells containing at least 20 respondents, the cells were combined across race classifications, and only then in a very limited way. To avoid combining certain race categories, a few weighting classes contained less than 20 respondents. This procedure of defining weighting classes led, of course, to different combinations of collapsed cells from one major stratum to the next.

Once the weighting classes were defined, pupil weights were computed. Following is the expression used to compute the weight, W_{Mijk}, for the k^{th} respondent contained in the i^{th} school and j^{th} weighting class in major stratum M for the Restandardization sample:

$$W_{Mijk} = W'_{Mi}\left(\frac{\sum_{i}^{T'_M} W'_{Mi}\, N_{Mij}}{\sum_{i} W'_{Mi}\, N'_{Mij}}\right), \qquad (3\text{-}10)$$

where

N_{Mij} = the number of pupils contained in the i^{th} participating school selected from major stratum M who belong to the j^{th} weighting class, and

N'_{Mij} = the number of the N_{Mij} pupils selected from the i^{th} school and the j^{th} weighting class who provided usable test data.

The pupil nonresponse factor that appears in equation 3-10 above is simply the sum of the school weights of all those pupils selected and contained in the j^{th} weighting class divided by the same sum for those pupils contained in the j^{th} weighting class and providing usable test data.

9.4 Pupil Weights for the Equating Study

The expression used to compute weights for the pupils participating in the Equating Study was composed of the expression used for the Restandardization weights given in equation 3-10 plus two additional factors. One factor reflects the random assignment of the classes to the testing sequences. Adding this factor to the weighting expression given in equation 3-10, the weight, W_{lMijk}, assigned to the k^{th} participating pupil contained in the i^{th} school and the j^{th} weighting class in major stratum M and belonging to an odd-numbered class, was the following:

$$W_{lMijk} = W'_{Mi}\left(\frac{C_{Mi}}{C_{lMi}}\right)\left(\frac{T'_M}{\sum_i W'_{Mi} N_{Mij}}\Big/\frac{T'_M}{\sum_i W'_{Mi} N'_{Mij}}\right), \qquad (3\text{-}11)$$

where

C_{Mi} = the number of classes contained in the i^{th} school participating from major stratum M, and

C_{lMi} = the number of classes that provided usable test data and administered the designated pair of tests (or test forms) in the sequence specified for the odd-numbered classes.

Likewise, the corresponding weight, W_{OMijk}, for the k^{th} participating pupil contained in the i^{th} school and the j^{th} weighting class in major stratum M and belonging to an even-numbered class was computed as follows:

$$W_{OMijk} = W'_{Mi}\left(\frac{C_{Mi}}{2\,C_{OMi}}\right)\left(\frac{T'_M}{\sum_i W'_{Mi}\,N_{Mij}} \Big/ \frac{T'_M}{\sum_i W'_{Mi}\,N'_{Mij}}\right), \qquad (3\text{-}12)$$

where

C_{OMi} = the number of classes that provided usable data and administered the designated pair of tests (or test forms) in the sequence specified for the even-numbered classes.

The factor "2," which appears in the denominator of the two previous weighting expressions, was inserted so that the sum of all pupils in the sample would be approximately equal to the total number of eligible pupils in the population, rather than twice this number.

The expected sum of the weights, across all 16 test matrices, of the pupils in the schools contained in a particular cell (or half-cell) is the same for each of the 56 cells (or half-cells). It turned out that the actual sums of weights for some of the 56 cells (or half-cells) varied substantially from the average cell sum. In fact, in one instance, the sum for the cell deviated from the average sum by about 50 percent. Therefore, one final adjustment was applied to the Equating weights in order to

equalize these cell sums. That is, the final weight, W'_{tMijk}, for a pupil contained in cell u was computed using the following expression:

$$W'_{tMijk} = W_{tMijk}(\bar{n}/n_u) , \qquad (3\text{-}13)$$

where

$$n_u = \sum_{\substack{\text{all students} \\ \text{in cell u}}} W_{tMijk} , \text{ and}$$

$$\bar{n} = (1/56) \sum_{u}^{56} n_u .$$

10. Survey Estimates

10.1 Introduction

The method used in the Equating Study to estimate the equating relationships was described in detail in the RFP and is not repeated here. However, the method used to estimate the variances of the equating relationships is given in Section 11 below.

In the Restandardization Study, estimates of percentiles and percentile ranks by grade and subtest (e.g., reading comprehension and vocabulary) were computed for the test that served as the "Anchor Test." Based on these estimates and the estimated equating relationships derived in the Equating Study, estimated norms were obtained for all seven tests used in the study.

The expressions used to make these estimates are summarized below. Also given below is a summary of the expressions that are appropriate to estimate the pupil mean for a particular grade and subtest and those appropriate to estimate percentiles and percentile ranks for schools. All the estimates discussed below are described in more detail in Appendix F, along with a specification of how to compute corresponding variance estimates.

10.2 Estimates of Pupil Percentiles and Percentile Ranks

The percentile rank corresponding to a particular test score, T, is the proportion, P, of eligible pupils in the population who have (expected) scores less than T. The P^{th} percentile for a test is that score, T_P, such that the proportion of all eligible pupils who have (expected) scores less than T_P is P. That is, T_P is the P^{th} percentile if

$$P = X/Y \ , \qquad (3\text{-}14)$$

where

Y = the total number of eligible pupils in the population (in a particular grade), and

X = the number of pupils in the specific grade who have (expected) scores less than T.

Also, P is the percentile rank for the score T_P.

The sample estimate, p, of the percentile rank, P, corresponding to a particular score, T, was computed as follows:

$$p = x'/y' , \tag{3-15}$$

where

y' is the sample estimate of Y, and

x' is the sample estimate of the number of pupils who have (expected) scores less than T.

The estimates y' and x' took on a simple form. The estimate y' was made by summing the weights of all pupils in the particular grade who provided usable test data. The estimate x' was made by summing the weights of those pupils who had test scores less than T.* That is, if W_i represents the weight assigned to the i^{th} respondent and A is the set of all pupils with scores less than T, the sample estimate, p, of the percentile rank for the score T was computed as follows:

$$p = x'/y' = \sum_{\substack{\text{all pupils} \\ \text{in A}}} W_i \Big/ \sum_{\substack{\text{all pupils} \\ \text{in sample}}} W_i . \tag{3-16}$$

---

* This summary of the procedure used to estimate percentile ranks is oversimplified. In particular, it does not take into account the fact that the discrete test score distribution was smoothed and approximated by a continuous one to estimate mid-percentile ranks, which were actually computed for the Restandardization Study norms.

The sample estimate of the P^{th} percentile T_p, was taken to be that score, t_p, such that

$$p = x'/y' , \tag{3-17}$$

where

y' is the sample estimate of Y used in equation 3-16, and

x' is the sample estimate of the number of pupils who have (expected) scores less than t_p.

10.3 The Estimate of the Pupil Mean

The mean score on a test, $\bar{X}$, for all pupils in the population in a particular grade is simply the total of all the (expected) scores of all eligible pupils, X, divided by the total number of eligible pupils, Y, in the population. That is,

$$\bar{X} = X/Y \tag{3-18}$$

The form of the expression for $\bar{X}$ is the same as that for population percentile rank, P, given in equation 3-14 in the previous subsection.

The sample estimate, $\bar{x}$, of a test mean, $\bar{X}$, is made by taking the ratio of the estimated total of all pupil scores, x', to the estimated total number of eligible pupils, y'. The estimate y' should be made in precisely the same way it was for estimating percentile ranks. The estimate x' should

be made by summing the products of the pupil weights and test scores over all pupils in the particular grade in the sample. That is, if W_i and x_i represent the weight and test score for the i^{th} pupil in the particular grade in the sample, the sample mean should be estimated as follows:

$$\bar{x} = x'/y' = \sum_i W_i x_i / \sum_i W_i . \qquad (3\text{-}19)$$

10.4 Estimation of School Percentiles and Percentile Ranks

Percentile ranks and percentiles for schools for a particular test are defined in a way similar to those for pupils. That is, T_P is the p^{th} percentile for schools for a particular grade and test if

$$P = X/Y , \qquad (3\text{-}20)$$

where

Y = the total number of (eligible) schools in the population that contain the particular grade, and

X = the number of schools having mean scores less than T_p.

Correspondingly, P is the percentile rank for the score T_P.

The expressions that should be used to estimate the school percentiles and percentile ranks are very similar to those used for pupils. The estimate, p, of a school percentile rank corresponding to test score T should be computed as the estimated total number of schools in the population containing the particular grade and having mean scores less than T divided by the

estimated total number of schools in the population. That is, if W_i represents the weight of the i^{th} school in the sample, A represents the set of schools in the sample with mean scores (for the particular grade) less than T, and B represents the set of all schools in the sample that contain the particular grade, the estimate, p, of the school percentile rank for score T should be made as follows:

$$p = x'/y' = \sum_{\substack{\text{all schools} \\ \text{in A}}} W_i \Big/ \sum_{\substack{\text{all schools} \\ \text{in B}}} W_i . \qquad (3\text{-}21)$$

From the estimated school percentile ranks, school percentiles are estimated in the same way that pupil percentiles were estimated from pupil percentile ranks.

11. Variance Estimation

11.1 Brief Summary of the Variance Estimation Procedure for the Restandardization Study

Variance estimates for the Restandardization Study are given in detail in Appendix E. Basically, since the survey estimates for the Restandardization Study discussed in the previous section are all ratio estimates, the variances of these estimates can be estimated using the

Taylor approximation to the variance of a ratio. That is, for an estimate, r, of the form x'/y', the variance can be estimated as follows:

$$\hat{\sigma}_r^2 = r^2 \left(\frac{\hat{\sigma}_{x'}^2}{(x')^2} + \frac{\hat{\sigma}_{y'}^2}{(y')^2} - 2 \frac{\hat{\sigma}_{x'y'}}{x'y'} \right) , \qquad (3\text{-}22)$$

where

$\hat{\sigma}_{x'}^2$ = the estimated variance of the estimated total, x',

$\hat{\sigma}_{y'}^2$ = the estimated variance of the estimated total, y', and

$\hat{\sigma}_{x'y'}$ = the estimated covariance between the estimated totals, x' and y'.

The estimates of the variances of the totals, x' and y', and the covariance between x' and y' are based, in general, on the variation between the sample estimates of the two schools selected from each final stratum. That is, if x'_{h1} and x'_{h2} represent the estimated totals (i.e., sums of the products of the weights and the corresponding values of the variable) for the two schools selected from major stratum h, the major contribution to the estimated variance of x' is computed as follows:

$$\hat{\sigma}_{x'}^2 = \sum_h^{L'} (x'_{h1} - x'_{h2})^2 , \qquad (3\text{-}23)$$

where

L' = the number of final strata that contained two participating schools.

There is a small additional component of the estimated variance of a total computed from the supplementary sample of missed schools. This component is based on the variation between estimates obtained from the two (or more) districts selected from each of the 82 district strata. Appendix E shows precisely how this component can be computed.

The estimate $\hat{\sigma}^2_{y'}$, is computed in the same way that $\hat{\sigma}^2_{x'}$, is computed, with y'_{h1} and y'_{h2} defined the same way that x'_{h1} and x'_{h2} are defined. The estimated covariance, $\hat{\sigma}_{x'y'}$, is computed in an analogous way as follows:

$$\hat{\sigma}_{x'y'} = \sum_{h}^{L'} (x'_{h1} - x'_{h2})(y'_{h1} - y'_{h2}) \ . \qquad (3\text{-}24)$$

This estimate also contains a small additional component computed from the supplementary sample of missed schools.

11.2 Estimation of the Variances for the Equating Study

The equating of scores of seven tests was involved in the Equating Study. Let W represent the test that served as the "Anchor Test" in the equating procedures. Also, let X represent any of the other tests, and let $W_{X_i}^*$ represent the converted score in terms of test W of the i^{th} score on test X. As indicated previously, the variance of the equated score, W_{X_i},* was estimated using McCarthy's method of balanced half-sample replication (BHSR).** Basically, the variance of an equated score, W_{X_i},*

** For a detailed discussion of this technique, see the reference given earlier in Section 8.2.

was estimated using this method by computing the variance among estimated equated scores based on a set of eight balanced half-samples. This is described in more detail in the following paragraphs.

The set of eight balanced half-samples: As mentioned previously, the Equating sample was partitioned into eight principal strata, each consisting of two groups of 28 schools. Each of the total of 16 groups of 28 schools filled out the 56 cells or half-cells of a test assignment matrix.

A half-sample for the Equating Study consists of eight of the 16 groups of 28 schools--one group from each of the eight principal strata. A particular half-sample can be represented by a sequence of eight signs, each sign being either a "+" or a "-." A plus sign indicates that the first member of a principal stratum is a member of the half-sample, whereas a minus sign indicates that the second member of a stratum is a member of the half-sample. For example, the following is the designation of a particular half-sample:

+ + - + - - + +

This half-sample contains the first members of principal strata 1, 2, 4, 7, and 8, and the second members of strata 3, 5, and 6.

In general, a set of balanced half-samples (BHSs) is determined by an orthogonal matrix of plus and minus signs. The order of such a matrix is k x k, where k must be a multiple of four. Each row of an orthogonal matrix determines a BHS, and each column designates a stratum. Since the

order of the matrix must be a multiple of four, the smallest orthogonal matrix that can be used to define a set of BHSs for the Equating Study is an 8 x 8 orthogonal matrix. The matrix that was used for this purpose is given in Table 3-8 below.

Table 3-8 The 8 x 8 Orthogonal Matrix Used to Define the Set of Eight Balanced Half Samples*

	Stratum Number							
BHS Number	1	2	3	4	5	6	7	8
1	+	-	-	+	-	+	+	-
2	+	+	-	-	+	-	+	-
3	+	+	+	-	-	+	-	-
4	-	+	+	+	-	-	+	-
5	+	-	+	+	+	-	-	-
6	-	+	-	+	+	+		
7	-	-	+	-	+	+	+	-
8	-	-	-	-	-	-	-	+

Again, a "+" in the above matrix indicates that the first member (group of 28 schools) in the corresponding principal stratum is included in the designated BHS, while a "-" indicates that the second member is included.

* This orthogonal matrix is determined by the first column. All other columns except the last one are generated from the previous column. That is, the second column is derived from the first column by lowering the first six signs in the first column by one row and moving the seventh sign to the first row. The last sign in the column remains fixed. Each of the third through the seventh columns is generated in the same way from the previous column. The signs contained in the last row and column of the matrix are all minus and they are not part of the generating process.

The "balance" in the set of BHSs comes from the orthogonality of the columns, not from the orthogonality of the rows. This balance implies that, for a linear estimate, the average of squared deviations of the BHS estimates from the entire-sample estimate provides an unbiased estimate of the variance of the entire-sample estimate. Unfortunately, it is a very difficult analytical problem to determine to what extent this property carries over to nonlinear estimation.

However, there is evidence that indicates that the BHSR procedure does produce good estimates in nonlinear cases. For example, McCarthy has compared variance estimates using the BHSR approach with those based on independent replications.** This comparison was made for ratio and regression estimates using data from the Health Examination Survey. The variance estimates based on the BHSR method were very close to the variance estimates based on independent replications, especially for ratio estimates.

The variance estimation expression: The standard BHSR variance estimation procedure is a direct extension of the procedure, mentioned earlier, that is known to have good properties in the linear case. The resulting estimate of the variance of an estimated equated score, $W_{X_i}^*$, was computed as follows:

$$\hat{\sigma}^2_{WX_i^*} = \frac{1}{8} \sum_j (W_{X_i}^{*(j)} - W_{X_i}^*)^2 , \qquad (3\text{-}25)$$

** See Philip J. McCarthy, Report on Development of Methods for Statistical Analysis of Data (Contract No. PH 86-65-94), New York State School of Industrial and Labor Relations, Cornell University, January 5, 1967, pp. 5-24, especially Tables 3 and 4, pp. 19-24.

where $W_{X_i}^{*(j)}$ denotes the estimated (converted) score on test W corresponding to score X_i, based on the j^{th} half-sample.

Separate weight adjustments for each BHS: Ideally, all weighting and estimation procedures that were carried out for the entire sample would have been carried out separately for each of the eight BHSs in computing equated scores. However, since both school and pupil response were high, there would probably be only a trivial difference in the nonresponse adjustments based on the entire sample and those based on a half-sample. Therefore, since determining weighting classes for pupil adjustments involves a large amount of time and effort, pupil nonresponse adjustments were not made separately for each half-sample.

However, school nonresponse adjustments were made separately for each half-sample. Also, the weight adjustments described in Section 9.4 used to equalize the sum of the weights for each of the 56 cells or half-cells were made separately for each half-sample.

CHAPTER 4

METHODOLOGY

1. Restandardization Study

The Restandardization Study required the administration of only the Anchor Test (MAT) to all students in the 4th, 5th, and 6th grades in a national sample of 918 schools with approximately 68,000 pupils per grade. Individual norms and school mean norms were developed at each of grades 4, 5, and 6, for all seven tests used in the Equating Study. Since observed score distributions were available only for the MAT, the statistical procedures for developing norms for the other six reading tests differed from those for the MAT. For each of the six non-MAT tests, norms were derived using the equating results and the MAT test score data from the Restandardization Study.

The detailed description of the restandardization procedures, applied identically to vocabulary, reading comprehension, and total reading tests at each of grades 4, 5, and 6, appears below.

1.1 Weighting Procedure

Each student in the Restandardization Study was assigned a weight which was thereafter associated with that student whenever his data were included in any analysis. The weight consisted of the product of two factors, each determined in the manner described below. (Computational formulas are presented in Section 9 of Chapter 3.)

Factor 1 was a school weight determined as the inverse of the school selection probability. An adjustment for school non-response was made within major stratum by multiplying the school weight for each remaining

school in the stratum by the ratio of the sum of the school weights for all schools selected from the stratum to the sum of the school weights for all schools in the stratum which actually participated in the study.

Factor 2 was a weight determined by membership in a particular "weighting class" defined by sex, race, and IQ level within each major stratum. That is, for each grade separately, all eligible students in the participating schools in a given major stratum were classified in a trivariate (sex x race x IQ level) frequency distribution as either usable or non-usable cases. (For details of the classification procedure and definitions of eligibility and usable cases, see Appendix F.) Each student carried his school weight in the frequency distribution. For each weighting class (i.e., for each cell in the trivariate layout), the second weighting factor was computed as the ratio of total number of cases to usable number of cases.

In order to avoid erratic weighting factors in weighting classes with small numbers of students, the trivariate distributions were examined with entries unweighted by school weights prior to the computation of factor 2. Generally, when the usable N was less than 20, cell frequencies were pooled. The guidelines for determining which cells to pool were generally as follows.

(a) IQ level was the first choice for collapsing: those with no IQ information were pooled with the average IQ level; extremes were pooled toward the average.

(b) On the relatively few occasions when collapse on IQ level was not effective, race served as the pooling variable: those with no race information were pooled with White and Other; Orientals were pooled with White; American Indians were pooled with either Blacks or Spanish-surnamed; Blacks and Spanish-surnamed were combined;

occasionally, all three latter groups were pooled.

1.2 MAT Individual Norms

For each of the two MAT subtests and for the total MAT test, at each of grades 4, 5, and 6, frequency distributions were obtained and smoothed by the Tukey-Cureton procedure. (For details of the smoothing procedure see Appendix G.) For each raw score point of each of the nine resulting smoothed distributions, mid-percentile ranks were developed and stanines assigned according to Table 4-1 below

Table 4-1. Range of Percentile Ranks Defining the Stanine Scale

Percentile Ranks Range	Corresponding Stanine
96.00 and above	9
89.45 - 95.99	8
77.35 - 89.44	7
59.88 - 77.34	6
40.14 - 59.87	5
22.67 - 40.13	4
10.57 - 22.66	3
4.02 - 10.56	2
4.01 and below	1

This resulted in three MAT individual norms tables, one for each grade, each including norms for vocabulary, reading comprehension, and total reading, which appear in Volume I of this report as Tables 577, 578, and 579.

1.3 Individual Norms for CAT, CTBS, ITBS, STEP II, SRA, and SAT

Each of the six non-MAT tests was equated to the MAT using eight procedure/method combinations. However, individual norms for each of these tests were generated using the MAT individual score norms and the equating tables for only four procedure/method combinations: Procedure 3 and the Average of Procedures 1 and 2; linear and equipercentile methods.

To obtain midpercentile ranks for (non-integer) MAT-equivalent scores, for each MAT individual norms table, midpercentile ranks were obtained (by linear interpolation) for MAT scores at each tenth of the interval between adjacent raw score points, for the entire range of the MAT raw score scale. Stanines were then associated with each of these scores through the interpolated value of the midpercentile rank and Table 4-1. This resulted in nine (vocabulary, reading comprehension, and total reading tests at each grade) "expanded" individual norms tables, with a mid-percentile rank and stanine corresponding to each tenth of a raw score point on the MAT.

Each raw score point on a given non-MAT test was then mapped into a midpercentile rank and stanine through its MAT-equivalent score obtained by each of the four procedure/method combinations. That is, each raw score point of a given non-MAT test was converted to its MAT-equivalent derived by a particular procedure/method combination, and the mid-percentile rank and stanine for the MAT value from the "expanded" MAT individual norms table were perforce associated with the non-MAT raw score value. Any non-MAT raw score value whose MAT-equivalent fell outside the MAT raw score range was assigned a midpercentile rank of 99+ at the upper end of the scale and 1- at the lower end.

As a result, four individual norms tables, one for each procedure/method combination, were developed for each non-MAT vocabulary, reading comprehension, and total reading test at each of grades 4, 5, and 6. These tables appear in Volume I of this report as indicated in the reference table below.

Table 4-2. Reference Table for Individual Score Norms for the Non-MAT Tests

	Table Numbers		
Grade	4	5	6
Procedure/Method Combination			
Procedure 3, linear	505-510	529-534	553-558
Procedure 3, equipercentile	511-516	535-540	559-564
Average of Procedures 1 and 2, linear	517-522	541-546	565-570
Average of Procedures 1 and 2, equipercentile	523-528	547-552	571-576

1.4 School Mean Norms for the MAT

At each of grades 4, 5, and 6, for the MAT vocabulary, reading comprehension, and total reading tests, mean scores for each school in the restandardization sample were computed. The distribution of school means, each weighted by the appropriate school weight, was produced and smoothed by the Tukey-Cureton procedure. Midpercentile ranks and stanines were then computed for each of these nine distributions in the same manner as for the MAT individual norms, resulting in three tables, one for each grade, each containing school mean norms for vocabulary, reading comprehension, and total reading scores. These appear as Tables 652, 653, and 654 of Volume I of this report.

1.5 School Mean Norms for CAT, CTBS, ITBS, STEP II, SRA, SAT

For the six non-MAT tests, it was required to obtain school means by conversion of MAT scores to each of these tests, using the equating tables provided by each of the four procedure/method combinations. Since for the equipercentile method, the school mean of the students' converted MAT scores is not equal to the converted MAT school mean, it was necessary to transform each individual student's MAT score to its equivalent on each of the other six tests and then to compute the school mean of the converted scores. For convenience in data management, school means for the linear method were produced in the same way as were those for the equipercentile method. Each student's MAT vocabulary, reading comprehension, and total reading scores were thus converted to each of the other six tests by the equating tables resulting from each of the four procedure/method combinations, for a total of 72 conversions per student.

At each grade, the 72 distributions of school means of the converted vocabulary, reading comprehension, and total reading scores derived from each of the four procedure/method combinations were produced and smoothed by the Tukey-Cureton procedure. As with the MAT, the school mean in each distribution was weighted by the school sampling weight. Thereafter, midpercentile ranks and stanines were calculated for the non-MAT tests in the same way as they were for the MAT.

For each of the six non-MAT tests at each grade, four derived norms tables, one from each procedure/method combination, each containing school mean norms for vocabulary, reading comprehension, and total reading scores, were thus produced. These tables appear in Volume I of this report as indicated in the reference table below.

Table 4-3. Reference Table for School Mean Norms for the Non-MAT Tests

	Table Numbers		
Grade	4	5	6
Procedure/Method Combination			
Procedure 3, linear	580-585	604-609	628-633
Procedure 3, equipercentile	586-591	610-615	634-639
Average of Procedures 1 and 2, linear	592-597	616-621	640-645
Average of Procedures 1 and 2, equipercentile	598-603	622-627	646-651

2. Equating Study

The design of the Equating Study required the administration of the seven tests (CAT, CTBS, ITBS, MAT, STEP II, SRA, SAT) in each test pair combination for each order of administration at each of grades 4, 5, and 6. The design is illustrated by Table 4-4, in which the row designates the test administered first and the column, the test administered second. It should be noted that, where a test was paired with itself, students actually took alternate forms of the test (the secondary form is designated by a †) rather than being retested with the same form. Any given off-diagonal cell thus represents one order of administration for a particular test-pair while its symmetric opposite represents the reverse order for that same test-pair. The counterbalanced order of administration for the alternate forms of the same test is represented by the double-diagonal cells. A given school-unit* was assigned to a particular

* A school-unit is defined to consist of a minimum of two classes at each of grades 4, 5, and 6. When an individual school did not fulfill this requirement, two or more schools were combined to form a "pseudo school", which then served as a school-unit. For details on the formation of pseudo schools, see Section 6 of Chapter 3.

Table 4-4. Design for Equating Study Test Administration

X_k (Test Administered Second)

ROWS:	TESTS:	COLUMNS: 1	2	3	4	5	6	7	COMMON TEST:
		CAT (2)	CTBS (2)	ITBS (2)	MAT (2)	STEP II (2)	SRA (2)	SAT (2)	
1	CAT (1)	CAT (1) CAT† (2) CAT† (1) CAT (2)	CAT (1) CTBS (2)	CAT (1) ITBS (2)	CAT (1) MAT (2)	CAT (1) STEP II (2)	CAT (1) SRA (2)	CAT (1) SAT (2)	CAT (1)
2	CTBS (1)	CTBS (1) CAT (2)	CTBS (1) CTBS† (2) CTBS† (1) CTBS (2)	CTBS (1) ITBS (2)	CTBS (1) MAT (2)	CTBS (1) STEP II (2)	CTBS (1) SRA (2)	CTBS (1) SAT (2)	CTBS (1)
3	ITBS (1)	ITBS (1) CAT (2)	ITBS (1) CTBS (2)	ITBS (1) ITBS† (2) ITBS† (1) ITBS (2)	ITBS (1) MAT (2)	ITBS (1) STEP II (2)	ITBS (1) SRA (2)	ITBS (1) SAT (2)	ITBS (1)
4	MAT (1)	MAT (1) CAT (2)	MAT (1) CTBS (2)	MAT (1) ITBS (2)	MAT (1) MAT† (2) MAT† (1) MAT (2)	MAT (1) STEP II (2)	MAT (1) SRA (2)	MAT (1) SAT (2)	MAT (1)
5	STEP II (1)	STEP II (1) CAT (2)	STEP II (1) CTBS (2)	STEP II (1) ITBS (2)	STEP II (1) MAT (2)	STEP II (1) STEP II† (2) STEP II† (1) STEP II (2)	STEP II (1) SRA (2)	STEP II (1) SAT (2)	STEP II (1)
6	SRA (1)	SRA (1) CAT (2)	SRA (1) CTBS (2)	SRA (1) ITBS (2)	SRA (1) MAT (2)	SRA (1) STEP II (2)	SRA (2) SRA† (2) SRA† (1) SRA (2)	SRA (1) SAT (2)	SRA (1)
7	SAT (1)	SAT (1) CAT (2)	SAT (1) CTBS (2)	SAT (1) ITBS (2)	SAT (1) MAT (2)	SAT (1) STEP II (2)	SAT (1) SRA (2)	SAT (1) SAT† (2) SAT† (1) SAT (2)	SAT (1)
COMMON TEST:		CAT (2)	CTBS (2)	ITBS (2)	MAT (2)	STEP II (2)	SRA (2)	SAT (2)	

X_j (Test Administered First)

NOTE: (1) indicates test taken as a first test.

(2) indicates test taken as a second test.

Table 4-5. General Design for Equating Study

X_k (Test Administered Second)

ROWS:	TESTS:	COLUMNS: 1 CAT (2)	2 CTBS (2)	3 ITBS (2)	4 MAT (2)	5 STEP II (2)	6 SRA (2)	7 SAT (2)	COMMON TEST:
1	CAT (1)	CAT (1) / CAT (2)	CAT (1) / CTBS (2)	CAT (1) / ITBS (2)	CAT (1) / MAT (2)	CAT (1) / STEP II (2)	CAT (1) / SRA (2)	CAT (1) / SAT (2)	CAT (1)
2	CTBS (1)	CTBS (1) / CAT (2)	CTBS (1) / CTBS (2)	CTBS (1) / ITBS (2)	CTBS (1) / MAT (2)	CTBS (1) / STEP II (2)	CTBS (1) / SRA (2)	CTBS (1) / SAT (2)	CTBS (1)
3	ITBS (1)	ITBS (1) / CAT (2)	ITBS (1) / CTBS (2)	ITBS (1) / ITBS (2)	ITBS (1) / MAT (2)	ITBS (1) / STEP II (2)	ITBS (1) / SRA (2)	ITBS (1) / SAT (2)	ITBS (1)
4	MAT (1)	MAT (1) / CAT (2)	MAT (1) / CTBS (2)	MAT (1) / ITBS (2)	MAT (1) / MAT (2)	MAT (1) / STEP II (2)	MAT (1) / SRA (2)	MAT (1) / SAT (2)	MAT (1)
5	STEP II (1)	STEP II (1) / CAT (2)	STEP II (1) / CTBS (2)	STEP II (1) / ITBS (2)	STEP II (1) / MAT (2)	STEP II (1) / STEP II (2)	STEP II (1) / SRA (2)	STEP II (1) / SAT (2)	STEP II (1)
6	SRA (1)	SRA (1) / CAT (2)	SRA (1) / CTBS (2)	SRA (1) / ITBS (2)	SRA (1) / MAT (2)	SRA (1) / STEP II (2)	SRA (1) / SRA (2)	SRA (1) / SAT (2)	SRA (1)
7	SAT (1)	SAT (1) / CAT (2)	SAT (1) / CTBS (2)	SAT (1) / ITBS (2)	SAT (1) / MAT (2)	SAT (1) / STEP II (2)	SAT (1) / SRA (2)	SAT (1) / SAT (2)	SAT (1)
COMMON TEST:		CAT (2)	CTBS (2)	ITBS (2)	MAT (2)	STEP II (2)	SRA (2)	SAT (2)	

X_j (Test Administered First)

NOTE: (1) indicates test taken as a first test.

(2) indicates test taken as a second test.

test-pair and thus tested half its classes at each grade in one order of administration and the other half in the opposite order.

The assignment of school-units to the cells of Table 4-4 was accomplished in the following manner. (A detailed description of the test assignment is presented in Section 8 of Chapter 3.) On the basis of socio-economic status, the total sample of 448 school-units was divided into eight principal strata, each consisting of 56 school-units, two from each of 28 final strata. This was followed by a balanced random splitting of the final stratum pairs, resulting in two sets of 28 school-units each for every principal stratum. Within each of the 16 sets so produced, the 28 school-units were then assigned randomly to the test-pairs represented in Table 4-4, and classes within each grade of each school-unit were assigned alternately to the corresponding upper or lower triangular cell, i.e., to each of the two orders of administration.

After all school-units and classes at each grade had been assigned to the appropriate cell of Table 4-4, the administration of the test-pairs in a given sequence yielded for each tested student six scores: vocabulary, reading comprehension, and total reading, for both the test taken first and the test taken second. Since for equating purposes, only the primary form of each test was used, the general design for the ensuing description of the Equating Study appears as Table 4-5, which is identical to Table 4-4 except that the secondary forms have been eliminated from the double diagonal cells. Note that two different groups of students were involved in the two administrations of the same test (double diagonal cells) whereas the same group of students was involved in the administration of a combination of two different tests (off-diagonal cells). This fact accounts for the differential treatment of double diagonals and off-diagonals throughout the Equating Study.

The objective of the Equating Study was to equate each of the subtest and total reading scores for a given test to the corresponding subtest and total scores of every other test for each of grades 4, 5, and 6, resulting in nine analogues of the general design, all treated identically with respect to the equating process. Each of the nine applications of the equating process produced eight sets of equating tables, one for each of the eight equating procedure/method combinations described below.

The equating process, for all analogues of the general design, consisted of both equipercentile and linear equating methods applied through each of three equating procedures. The three equating procedures represented different ways of organizing and utilizing the data available to derive the requisite test score distributions (equipercentile method), and means and standard deviations (linear method) for each test. From Table 4-5, it is evident that approximately 1/7 of the equating sample took the test common to a given row as a first test while an entirely nonoverlapping 1/7 of the equating sample took the same test, common to the corresponding column, as a second test. It was thus possible to develop frequency distributions for the equipercentile method and equating parameters for the linear method from three different data bases.

Procedure 1 utilized data from the seven tests when administered as second tests to estimate distributions, means, and variances for the entire sample to equate these tests when administered as first tests. The common data base for equating the tests when administered as first tests was generated in two stages. The first consisted of a preliminary equating to the MAT of each of the six non-MAT tests when taken as first tests. For the preliminary equating, either a frequency distribution (equipercentile method) or a mean

and variance (linear method) for each off-diagonal cell of Table 4-5 was estimated from the test taken second in its respective column and then collapsed over columns to produce a single estimated frequency distribution or mean and variance for each test when taken as a first test (row). The preliminary equating tables were then used to convert the non-MAT test scores when taken as second tests to the raw score scale of the MAT, resulting in a common score scale for all students in the equating sample. The preliminary equating was then followed by an iterated operational equating procedure which estimated the frequency distribution or mean and variance for each test common to a given row directly from the MAT-equivalent scores for the tests when taken as second tests for the entire sample. The iterations were continued until convergence indicated stabilization of the estimated distributions and moments or to a maximum of four. The set of seven distributions or seven means and standard deviations resulting from the final iteration of the operational equating procedure for each of the two equating methods was then used to equate each test to every other test, yielding the final equating tables for Procedure 1.

Procedure 2 was just the reverse, utilizing test data from the first administration to estimate distributions and parameters for the entire sample for equating the tests when administered as second tests. For Procedure 2, the cell distributions (equipercentile method) and means and variances (linear method) used in the preliminary equating of each test to the MAT were estimated from the test taken as a first test in their respective rows and then collapsed over rows, yielding a single estimated frequency distribution or mean and variance for each test when administered as a second test (column). The operational equating procedure, analogously, estimated the frequency distribution

or mean and variance for each column directly from the MAT-equivalent scores for the tests when taken as first tests for the entire sample. The final equating tables for Procedure 2 were generated from the seven distributions (equipercentile method) or the seven sets of means and standard deviations (linear method) of the test scores when administered second produced by the final operational equating iterations for Procedure 2.

The final equating tables from Procedures 1 and 2 were then averaged for both the linear and equipercentile methods to average the order of administration factor explicit in the manner in which they were generated.

Procedure 3 differed from Procedures 1 and 2 in two important respects. First, no estimation procedure was used to generate the frequency distributions (equipercentile method) or the means and variances (linear method), and, second, corresponding row and column test data were combined for each test, regardless of order of administration. Thus, each of the seven distributions and means and standard deviations yielded by Procedure 3 represented only the 2/7 of the total equating sample which actually took the test in question rather than the entire sample. As a result, each set of equating tables obtained by Procedure 3 was based on seven overlapping but not identical samples, was in no sense mediated by one or more prior equatings to the MAT, and was not characterized by a particular order of administration.

The detailed operations constituting each of the equating procedures are described below.

2.1 Weighting Procedure

Each student in the Equating Study was assigned a weight which was thereafter associated with that student whenever his data were included in any

analysis. The weight consisted of the product of four factors, each determined in the manner described below. (Computational formulas and weighting rationale are presented in section 9 of Chapter 3.)

Factors 1 and 2, the school weight and the individual weight determined by membership in a particular sex x race x IQ category, respectively, were computed in the same manner as for the Restandardization Study (See Appendix F).

Factor 3 was an adjustment for equalizing the probability that a class would be assigned to each of the two orders of administration. The adjustment was applied separately for each order of administration, and consisted of the ratio of half the total number of classes in the school-unit to the number of classes actually tested in the given order of administration.

Factor 4 was an adjustment to render all 56 cells of Table 4-5 equal in size. The adjustment was made by the ratio of the average weighted number of students per cell to the weighted number for the given cell.

For each student, then,the final operational weight used throughout the main and all special subgroup analyses was the product of factors 1, 2, 3, and 4. In each of the eight replications used to estimate the equating error variance, however, factor 4 was recomputed using only the students included in the given replication. Thus, the final operational weight for each student in the error variance estimation procedure was the product of factors 1, 2, 3, and a revised factor 4.

2.2 Notation

The notational conventions used throughout the description of the equating methodology will ordinarily be defined in the text or be evident from the discussion itself. In addition, a detailed and exhaustive presentation of all notation, organized by section, appears in Appendix H.

2.3 Equipercentile Method: Procedure 1

Preliminary Equating Trial: The first step in the preliminary equating (trial number t = 0) was to estimate the frequency distribution of the test taken first, X_j, in each cell of Table 4-5 using the marginal distribution of the test taken second, X_k, in the appropriate column and the bivariate distribution of X_j and X_k in the given cell. It was thus required to obtain, for each column k of Table 4-5, the frequency distribution of X_k for all students in that column (including those in the lower but not the upper diagonal cell), resulting in seven column marginal frequency distributions, one for each test taken second, each based on approximately 1/7 of the total equating sample at a given grade. For all 42 off-diagonal cells, the bivariate distributions of X_j (ordinate) vs X_k (abscissa) were generated.

Each cell bivariate distribution and the corresponding column distribution of X_k were input to the frequency estimation procedure (details of which are presented in Appendix J) by means of which the frequencies in each array in the cell bivariate distribution were expanded to estimate the frequency distribution expected if all students in column k had taken X_j as a first test. From each cell bivariate distribution, then, a marginal distribution of X_j, estimated for about 1/7 of the equating sample, was obtained.

This process resulted in a set of six estimated marginal distributions for each X_j (i.e., in each row), which were summed to produce a single estimated marginal distribution for X_j, based on approximately 6/7 of the total equating sample.

At this point, then, seven estimated marginal distributions were available, one for each X_j. Each was smoothed by the Tukey-Cureton procedure (for details, see Appendix G) and used to equate each test to the MAT by the equipercentile

method (for details of the equipercentile equating procedure, see Appendix K).

The preliminary equating thus finally resulted in a set of conversion tables, each expressing the raw score value of a given test X_j on the MAT scale to 1 decimal accuracy. The set of first test conversion tables produced by the preliminary (t=0) equating trial was designated $\{{}_0X_j^*\}$, with * indicating conversion to the MAT raw score scale.

Operational Equating Trials: Using the preliminary conversion tables, each score on the tests administered second, X_k, was expressed as its MAT-equivalent value, ${}_0X_k^*$. Each row contained the scores for a given first test X_j, and, as a result of the score conversion operation, the second test scores in each row were on the MAT raw score scale. Thus it was possible to estimate the marginal distribution of each test X_j for the entire equating sample by expanding the bivariate distribution of X_j with ${}_0X_k^*$ in a given row in accordance with the marginal distribution of the MAT converted scores over the entire table. The marginal distribution of each X_j obtained from the expanded bivariate distribution was smoothed using the Tukey-Cureton procedure, and the seven smoothed distributions so produced were used to equate each X_j to the MAT by the equipercentile equating method, yielding the set of conversion tables $\{{}_1X_j^*\}$ and concluding the first (t=1) operational equating trial.

For the next (t=2) operational trial, each second test score X_k was converted to its MAT-equivalent, ${}_1X_k^*$, using the conversion tables resulting from the first operational equating trial rather than those from the preliminary equating trial. Thereafter, the second operational trial proceeded like the first. Marginal distributions for each X_j estimated from the entire table were smoothed and equated to the MAT by the equipercentile method, yielding another set of MAT conversion tables, $\{{}_2X_j^*\}$.

The conversion tables from the first and second iterations of the operational equating, $\{{}_1X_j^*\}$ and $\{{}_2X_j^*\}$, were then compared for convergence within a difference of .1 at each raw score value of each test. If the criterion was satisfied, the iterative procedure terminated at this point and the seven smoothed frequency distributions which emerged from the second iteration were accepted as final.

If convergence was not achieved, a third iteration and, if necessary, a fourth, were performed. The sets of conversion tables produced by each iteration were used recursively to generate those for the next iteration. Whether or not convergence was achieved at the fourth iteration, the procedure was terminated and the smoothed distributions for each of the seven first tests resulting from the fourth iteration were considered final.

Equating Each Test to Every Other Test: The seven final smoothed distributions generated by the last operational trial were used to equate each test X_j to every other test $X_{j'}$, resulting in seven equating tables, one for each test to every other test, which are the final product of Procedure 1 for the equipercentile method. The seven tables for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-6. Reference Table for Equipercentile Method/Procedure 1 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	8- 14	176-182	344-350
Reading Comprehension	64- 70	232-238	400-406
Total Reading	120-126	288-294	456-462

2.4 Equipercentile Method: Procedure 2

Procedure 2 was identical to Procedure 1 except that the operations applied to tests taken second, X_k, in Procedure 1 were here applied analogously to tests taken first, X_j, and the equating of first tests X_j to $X_{j'}$ from Procedure 1 was replaced by the equating of second tests X_k to $X_{k'}$ in Procedure 2. Procedure 2 resulted in seven equating tables, one for each test to every other test, when administered as second tests. The seven tables for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-7. Reference Table for Equipercentile Method/Procedure 2 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	22- 28	190-196	358-364
Reading Comprehension	78- 84	246-252	415-420
Total Reading	134-140	302-308	470-476

2.5 Equipercentile Method: Average of Procedures 1 and 2

As a result of having carried out both Procedure 1 and Procedure 2, two equating tables were produced for each test equated to every other test. Each corresponding pair of tables, representing the result of equating the test administered as a first test and as a second test for Procedures 1 and 2, respectively, was averaged to produce a single table for each test to every other test.

The final product of the averaging of Procedures 1 and 2 was seven tables for which the effect of differential order of administration was averaged. The seven such tables for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-8. Reference Table for Equipercentile Method/Average of Procedures 1 and 2 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	50- 56	218-224	386-392
Reading Comprehension	106-112	274-280	442-448
Total Reading	162-168	330-336	498-504

2.6 Equipercentile Method: Procedure 3

Each of the seven distributions used for equating by Procedure 3 was obtained by pooling data for a given test in a corresponding row and column (including both of the double diagonal cells) and smoothing by the Tukey-Cureton procedure. The set of seven smoothed distributions so produced were then used to equate each test to every other test by the equipercentile method.

The seven final tables, one for each test equated to every other test, resulting from Procedure 3 for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-9. Reference Table for Equipercentile Method/Procedure 3 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	36- 42	204-210	372-378
Reading Comprehension	92- 98	260-266	428-434
Total Reading	148-154	316-322	484-490

2.7 Linear Method: Procedure 1

Preliminary equating: The first step in the preliminary equating trial was to estimate for each off-diagonal cell of Table 4-5, the mean and variance for each test when administered first, X_j, based on all students in the appropriate column k, (including the lower but not the upper diagonal cell) who took X_k as a second test. This required, for each off-diagonal cell, the computation of the mean and variance of X_j, designated $M_{j(k)}$ and $V_{j(k)}$, respectively, the mean and variance of X_k, designated $M_{k(j)}$ and $V_{k(j)}$, and the covariance of X_j and X_k, designated Cov(jk). In addition, for each column, the mean and variance of X_k, designated M_k and V_k, respectively, were calculated.*

Then the estimated means and variances for each of the 42 off-diagonal cells, designated $\hat{M}_{j(k)}$ and $\hat{V}_{j(k)}$, respectively, were obtained by the following formulas.**

$$(1) \qquad \hat{M}_{j(k)} = M_{j(k)} + \left[\frac{Cov(jk)}{V_{k(j)}}\right]\left[M_k - M_{k(j)}\right]$$

* Computational versions of all linear equating formulas, correspondingly numbered, appear in Appendix L.

** The rationale for these estimation procedures may be found in: Angoff, W. Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Ed. Washington, D.C.: American Council on Education, 1971. Pp. 508-600.

$$(2) \qquad \hat{V}_{j(k)} = V_{j(k)} + \left(\frac{Cov(jk)}{V_{k(j)}} \right)^2 \left[V_k - V_{k(j)} \right]$$

Equation (1) produced a matrix of $\hat{M}_{j(k)}$, each row of which contained six estimates of the mean of X_j, each obtained from the test taken as a second test in the appropriate column (except the column corresponding to the same test administered second - i.e., the diagonal cells were empty), and each representing approximately 1/7 of the equating sample. An analogous matrix of $\hat{V}_{j(k)}$ was generated by Equation (2).

For each row j of the matrix of $\hat{M}_{j(k)}$, the six estimates were combined to get a single estimate of the mean of X_j, designated ${}_0\hat{M}_j$ to indicate the value obtained from the preliminary (t=0) trial, based on approximately 6/7 of the equating sample. Each estimate was weighted by N_k, the number of students in column k.

$$(3) \qquad {}_0\hat{M}_j = \frac{\sum_{\substack{k=1 \\ k \neq j}}^{7} N_k \hat{M}_{j(k)}}{\sum_{\substack{k=1 \\ k \neq j}}^{7} N_k}$$

Similarly, the six estimates contained in each row of the matrix of $\hat{V}_{j(k)}$ were pooled to get a single estimate of the variance of X_j, designated ${}_0\hat{V}_j$, based on approximately 6/7 of the equating sample.

$$(4) \qquad {}_0\hat{V}_j = \frac{\sum_{\substack{k=1 \\ k \neq j}}^{7} (N_k - 1) \hat{V}_{j(k)} + \sum_{\substack{k=1 \\ k \neq j}}^{7} N_k \left(\hat{M}_{j(k)} - {}_0\hat{M}_j \right)^2}{\left(\sum_{\substack{k=1 \\ k \neq j}}^{7} N_k \right) - 1}$$

The seven values of $_0\hat{M}_j$ and $_0\hat{V}_j$ so obtained were then used to compute the linear equating parameters (to 4 decimal accuracy) for the preliminary equating of each non-MAT test X_j to the MAT, when taken as first tests.

$$(5) \qquad {}_0A_j = \left(\frac{{}_0\hat{V}_4}{{}_0\hat{V}_j} \right)^{1/2}$$

$$(6) \qquad {}_0B_j = {}_0\hat{M}_4 - {}_0A_j \, {}_0\hat{M}_j$$

The preliminary equating for each X_j to the MAT, resulting in the set of preliminary equating tables $\{{}_0X_j^*\}$, was carried out by application of the following formula to each raw score value of each test X_j.

$$(7) \qquad {}_0X_j^* = {}_0A_j \, X_j + {}_0B_j$$

Operational Equating Trials: For equipercentile equating, it was necessary as a first step to convert each score on X_k, taken as a second test, to its MAT-equivalent, ${}_0X_k^*$, derived by the preliminary equating, in order to provide a common score scale for the entire equating sample. For linear equating, however, the only required statistics were means, variances, and covariances of the converted scores, which could be obtained by direct application of the parameters to the original means, variances, and covariances. Hence, in lieu of and for equating purposes conceptually

equivalent to the score conversion operation carried out at this point in the equipercentile equating, the mean and variance of ${}_0X^*_k$ (implicitly) and the covariance of X_j and ${}_0X^*_k$ for each row j (excluding the diagonal cell) of Table 4-5, and designated, respectively, by ${}_0M^*_{(j-)}$, ${}_0V^*_{(j-)}$, and ${}_0Cov(j^*)_{(j-)}$, were obtained as follows.

Note: Since in the formulas below, the statistics converted to their MAT-equivalent values were those for the tests when administered second, it was notationally convenient to indicate the use of first test parameters for so doing by defining ${}_0A'_k \equiv {}_0A_j$ and ${}_0B'_k \equiv {}_0B_j$ for $j = k$.

(8)
$$ {}_0M^*_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} N_{k(j)} \left[{}_0A'_k M_{k(j)} + {}_0B'_k \right]}{N_{j(-)}} $$

(9)
$$ {}_0V^*_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} (N_{k(j)}-1)({}_0A'_k)^2 V_{k(j)} + \sum_{\substack{k=1 \\ k\neq j}}^{7} N_{k(j)} \left[{}_0A'_k M_{k(j)} + {}_0B'_k - {}_0M^*_{(j-)} \right]^2}{N_{j(-)}-1} $$

(10)
$$ {}_0Cov(j^*)_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} N_{k(j)} \left[{}_0A'_k Cov(jk) + {}_0A'_k M_{k(j)} M_{j(k)} + {}_0B'_k M_{j(k)} - M_{j(-)} {}_0M^*_{(j-)} \right]}{N_{j(-)}} $$

where $M_{j(-)}$ is the row mean of X_j, and $N_{j(-)}$, the row sample size, excluding the diagonal cell. $N_{k(j)}$ is the sample size in the j, $k^{\underline{th}}$ off-diagonal cell.

In addition, the mean and variance of the MAT-equivalent scores from the preliminary equating for tests when administered second, ${}_0X^*_k$, over all cells in Table 4-5, including the lower diagonals, were computed. They were designated ${}_0M^*$ and ${}_0V^*$, respectively.

$$(11) \qquad {}_0M^* = \frac{\sum_{k=1}^{7}\sum_{j=1}^{7} N_{k(j)}\left({}_0A'_k\, M_{k(j)} + {}_0B'_k \right)}{\sum_{k=1}^{7}\sum_{j=1}^{7} N_{k(j)}}$$

$$(12) \qquad {}_0V^* = \frac{\sum_{k=1}^{7}\sum_{j=1}^{7}\left[\left(N_{k(j)} - 1 \right)\left({}_0A'_k \right)^2 V_{k(j)} + N_{k(j)}\left({}_0A'_k\, M_{k(j)} + {}_0B'_k - {}_0M^* \right)^2 \right]}{\left(\sum_{j=1}^{7}\sum_{k=1}^{7} N_{k(j)} \right) - 1}$$

With the MAT-equivalent scores providing a common base, the total table converted-score statistics, ${}_0M^*$ and ${}_0V^*$, were used in conjunction with the ${}_0M^*_{(j-)}$, ${}_0V^*_{(j-)}$, and $Cov(j^*)_{(j-)}$ for each row j to estimate the mean and variance for each test, X_j, when administered as a first test, for the entire equating sample. These means and variances were designated ${}_1\hat{M}_j$ and ${}_1\hat{V}_j$, respectively, indicating the values obtained from the first (t=1) operational trial.

$$(13) \qquad {}_1\hat{M}_j = M_{j(-)} + \left[\frac{{}_0Cov(j^*)_{(j-)}}{{}_0V^*_{(j-)}} \right] \left[{}_0M^* - {}_0M^*_{(j-)} \right]$$

$$(14) \qquad {}_1\hat{V}_j = V_{j(-)} + \left[\frac{{}_0Cov(j^*)_{(j-)}}{{}_0V^*_{(j-)}} \right]^2 \left[{}_0V^* - {}_0V^*_{(j-)} \right]$$

With these first operational estimates of the mean and variance for each test X_j, the first operational parameters for equating each test to the MAT were computed.

$$(15) \qquad {}_1A_j = \left[\frac{{}_1\hat{V}_4}{{}_1\hat{V}_j} \right]^{1/2}$$

$$(16) \qquad {}_1B_j = {}_1\hat{M}_4 - {}_1A_j \; {}_1\hat{M}_j$$

The set of equating tables produced by application of the above parameters to each raw score value of each non-MAT test was designated $\{{}_1X^*_j\}$.

$$(17) \qquad {}_1X^*_j = {}_1A_j \; X_j + {}_1B_j$$

For the second (t=2) operational trial, the statistics for second test scores converted to MAT-equivalents were determined using the conversion parameters resulting from the first operational trial rather than those from the preliminary trial. That is, ${}_1A'_k$ and ${}_1B'_k$, defined analogously to ${}_0A'_k$ and ${}_0B'_k$, replaced these latter values in Equations (8) through (12), generating ${}_1M^*_{(j-)}$, ${}_1V^*_{(j-)}$, and ${}_1Cov(j^*)_{(j-)}$ for each row and ${}_1M^*$ and ${}_1V^*$ for the entire table. Equations (13) through (16) were then applied, incrementing t by 1 throughout, and resulting finally in a second operational set of equatings of each first test to the MAT, $\{{}_2X^*_j\}$.

The first two operational sets of equating tables, $\{{}_1X^*_j\}$ and $\{{}_2X^*_j\}$, were then compared for convergence within a difference of .1 at each raw score value of each test. If the criterion was satisfied, the iterative procedure was terminated at this point and ${}_2\hat{M}_j$ and ${}_2\hat{V}_j$ for each X_j were accepted as the final estimates for equating each test to every other test when all were administered as first tests.

If convergence was not achieved, the iterative procedure was carried out a third, and, if necessary, a fourth time, using Equations (8) through (17) with t incremented by 1 at each successive trial. Whether or not convergence was achieved by the fourth iteration, the procedure was terminated and the estimated means and variances, ${}_4\hat{M}_j$ and ${}_4\hat{V}_j$, for each test X_j were accepted as final.

Equating Each Test to Every Other Test: The seven final estimated means and variances, ${}_f\hat{M}_j$ and ${}_f\hat{V}_j$, for each test X_j were then used to compute linear parameters for equating each test to every other test $X_{j'}$, when all were administered as first tests.

$$(18) \qquad A_{j \to j'} = \left(\frac{\hat{f}V_{j'}}{\hat{f}V_{j}} \right)^{1/2}$$

$$(19) \qquad B_{j \to j'} = \hat{f}M_{j'} - A_{j \to j'} \, \hat{f}M_{j}$$

The converted scores, designated $X_{j \to j'}$, are determined by application of the following formula to each raw score value of each test.

$$(20) \qquad X_{j \to j'} = A_{j \to j'} \, X_{j} + B_{j \to j'}$$

The final product of Procedure 1 for the linear method of equating was seven equating tables, one for each test to every other test, for each of the nine applications of the general design, appearing in this report as indicated in the reference table below.

Table 4-10. Reference Table for Linear Method/Procedure 1 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	1- 7	169-175	337-343
Reading Comprehension	57- 63	225-231	393-399
Total Reading	113-119	281-287	449-455

2.8 Linear Method: Procedure 2

Procedure 2 was identical to Procedure 1, except that all j and k subscripts were interchanged as were all references to rows and columns, and upper and lower diagonal cells.

The final product of Procedure 2 was a set of seven equating tables, one for each test to every other test when all were taken as second tests, for each analogue of the general design. The set of seven tables for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-11. Reference Table for Linear Method/Procedure 2 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	15- 21	183-189	351-357
Reading Comprehension	71- 77	239-245	407-413
Total Reading	127-133	295-301	463-469

2.9 Linear Method: Average of Procedures 1 and 2

As a result of the application of both Procedure 1 and Procedure 2, two equating tables were generated for each test equated to every other test. Each corresponding pair of tables, representing the result of equating a given test when administered first and the same test when administered second for Procedures 1 and 2, respectively, was averaged to produce a single table for each test to every other test.

The final product of the averaging of Procedures 1 and 2 was seven tables for which the effect of differential order of administration was averaged.

The seven such tables for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-12. Reference Table for Linear Method/Average of Procedures 1 and 2 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	43- 49	211-217	379-385
Reading Comprehension	99-105	267-273	435-441
Total Reading	155-161	323-329	491-497

2.10 Linear Method: Procedure 3

The means and variances used for computing the Procedure 3 linear equating parameters were obtained by pooling corresponding row (test taken as a first test) and column (test taken as a second test) data for each test. Application of these parameters resulted in a set of seven equating tables, one for each test to every other test, for each of the nine analogues of the general design. These tables appear in this report as indicated in the reference table below.

Table 4-13. Reference Table for Linear Method/Procedure 3 Equating Results

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	29- 35	197-203	365-371
Reading Comprehension	85- 91	253-259	421-427
Total Reading	141-147	309-315	477-483

2.11 Estimation of Equating Error

In order to estimate empirically the error of equating associated with the MAT-equivalent scores for each non-MAT test, the equating operations (both linear and equipercentile) of Procedures 1, 2, 3, and the Average of Procedures 1 and 2 were replicated eight times on random half-samples of schools. (Details of the assignment of schools to each of the eight half-samples are presented in Section 11 of Chapter 3.)

For each random half-sample of schools, the six non-MAT tests were equated only to the MAT in the same way as in the main equating analysis (i.e., for the total sample) described in the preceding sections. Thus, the eight replications yielded eight sets of equating tables for each of the procedure/method combinations. The tables resulting from Procedures 1 and 2 were averaged for both linear and equipercentile methods. Estimates of equating error were thus determined for equating of each test to the MAT by Procedures 1, 2, 3, and the Average of Procedures 1 and 2, for both linear and equipercentile methods. The error in equating each non-MAT test to the MAT was defined as the square root of the average squared deviation of the MAT-equivalents determined by each of the eight replications from the corresponding value determined by the main analysis. Thus, for a given procedure/method combination, the equating error for each (i.e., the ℓ^{th}) raw score value of a given test X_j (or X_k), designated $\hat{E}(X^*_{\ell j})$, was calculated by the following formula, where r = the replication index, r = 1, ..., 8.

$$\hat{E}(X^*_{\ell j}) = \sqrt{\frac{\sum_{r=1}^{8} \left(X^*_{\ell j} - X^*_{\ell jr} \right)^2}{8}}$$

The eight tables of equating error estimates, one for each procedure/method combination, for each of the nine applications of the general design appear in this report as indicated in the reference table below.

Table 4-14. Reference Table for Estimates of Equating Error

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	2077-2084	2101-2108	2125-2132
Reading Comprehension	2085-2092	2109-2116	2133-2140
Total Reading	2093-2100	2117-2124	2141-2148

2.12 Conditional Error of Equating

The estimate of error determined in section 2.11 is entirely attributable to sampling variability in the equating process itself. The conditional error of equating X_j to $X_{j'}$, on the other hand, defined as the standard deviation of observed scores on X_j, around the equivalent score on $X_{j'}$, for each value of $X_{j'}$, represents several sources of error, including equating error, measurement error, etc.

Unlike all of the preceding analyses, the conditional error of equating was calculated only for a subset of the total equating sample and only for a subset of the total number of test-pairs. That is, for each of grades 4, 5, and 6, two distinct test-pair combinations were randomly chosen from the off-diagonal portion of Table 4-5. For each of these six test-pairs, the conditional error was determined for the equating of vocabulary, reading comprehension, and total reading scores by each of four procedure/method combinations: Procedure 3 and the Average of Procedures 1 and 2; linear

and equipercentile methods.

Since all four procedure/method combinations have, in some sense, averaged the effects of differential order of administration in the determination of the final equating tables, the two groups of students involved in the counterbalanced order of administration for each of the six test-pairs were combined for the computation of the conditional error. The direction of equating was taken to be from the test designated by the row to the test designated by the column in the upper triangular portion of Table 4-3. Because corresponding cells in the upper and lower triangular portions were combined for this analysis, the data for both the test equated from the test equated to were about equally divided between first and second administrations.

Each student, i, in a given pair of cells of Table 4-3 had observed scores $X_{i\ell j}$ and $Y_{imj'}$, representing the ℓ^{th} and m^{th} raw score values of X_j and $X_{j'}$, respectively. Associated with his raw score on test X_j was also an equivalent score, determined by a particular procedure/method combination, on test $X_{j'}$, designated $X_{i\ell,j\to j'}$ for the ℓ^{th} raw score value of X_j. Then the computation of the conditional error of equating for a given raw score value, $X_{\ell j}$, when converted to the raw score scale of $X_{j'}$, designated $CE(X_{\ell j\to j'})$, was computed as follows, where n_ℓ = the (weighted) number of students in the given pair of cells with raw score $X_{\ell j}$.

$$CE(X_{\ell j\to j'}) = \sqrt{\frac{\sum_{i=1}^{n_\ell} (X_{imj'} - X_{i\ell,j\to j'})^2}{n_\ell}}$$

A single estimate of the conditional error of equating from X_j to $X_{j'}$, averaged over all raw score values $X_{\ell j}$, was also obtained.

$$CE(X_{j \to j'}) = \sqrt{\frac{\sum_{\ell=0}^{\max(X_j)} \sum_{i=1}^{n_\ell} (X_{imj'} - X_{i\ell, j \to j'})^2}{\sum_{\ell=0}^{\max(X_j)} n_\ell}}$$

The conditional error of equating for each of the four procedure/method combinations for each test-pair for vocabulary, reading comprehension, and total reading appear in Volume 30 of this report as Tables 2170 through 2187.

2.13 Equating Analyses for Special Subgroups

In order to examine possible differences in equating results for students differing in sex, IQ, race, and socio-economic status, the total equating sample was divided into eleven special subgroups using the following bases for classification.

Sex was determined by the student's response on his answer sheet, resulting in the following breakdown of the total equating sample:

1. All boys
2. All girls

IQ was determined by the response to question 5 of the "For Test Administrator Use Only" section of the answer sheet for each student, resulting in the following breakdown:

3. High IQ: (above 110)

4. Medium IQ: (between 90 and 110)

5. Low IQ: (below 90)

Race was determined by the response to question 2 of the "For Test Administrator Use Only" section of the answer sheet for each student, resulting in the following breakdown:

6 Caucasian

7. Black

8. Spanish surname

Non respondents in any of the above categories were excluded from the special subgroup analyses.

Socioeconomic information was collected only at the school level through questions 2 (income distribution) and 3 (percentage of families on public welfare) on the Pretest Information Form. SES assignment was accomplished by classifying schools on the basis of an ad hoc examination of the responses to these two questions. The three levels so generated may be characterized as follows:

9. High SES: (1) at least 50% of families at income level $10,000 and over with no more than 20% below $6,000

(2) no more than 10% on public welfare

10. Medium SES: (1) more than 50% of families between $3,000 and $10,000 income level, subject to the restriction that such a school would not qualify for either the high or low level (i.e., with the remaining families split between the lowest and highest income classifications)

(2) no more than 50% on public welfare

11. Low SES: (1) more than 50% of families below $6,000 and not more than 10% above $10,000 income level

(2) no restriction on percentage on public welfare

The few schools for which no data were available were allocated to SES level according to major stratum sampling classification, conservatively biased toward medium SES level.

For each of these eleven special subgroups, at each grade, each test was equated to every other test by Procedure 3 only, using both linear and equipercentile equating methods, exactly as described for the total sample analyses. The results of these analyses appear in Volumes XI through XXI of this report, one for each of the eleven subgroups as sequenced here.

3. Correlations

Table 4-4 formed the basis for computation of alternate forms reliabilities for each of the seven tests as well as of the intercorrelations of each test with every other test. The procedure described herein was applied to each of the nine analogues of Table 4-4; that is, to the vocabulary, reading comprehension, and total reading scores at each of grades 4, 5, and 6.

For each cell of Table 4-4, the correlation between the two tests was computed,* yielding, for the off-diagonal cells, the set of observed inter-test correlations for each order of administration of test-pairs. For the double diagonal cells, this resulted in observed alternate forms reliabilities for each test for each order of administration of the two forms.

---

* Computational formulas appear in Appendix L.

The above correlations were then averaged over orders of administration utilizing Fisher's z-equivalents. That is, for each test-pair, the two correlations corresponding to the two orders of administration were transformed to z-equivalents; the z's, weighted by sample size, were averaged, and the average z was then expressed as an average r. In the resulting correlation matrices, averages of corresponding off-diagonal test-pair correlations are presented as the upper triangular off-diagonal elements and the average reliabilities appear as diagonal elements.

All observed off-diagonal correlations, both unaveraged and averaged, were then corrected for attenuation due to unreliability of the tests concerned. In all cases, the reliability used in the correction was the value obtained by averaging over the reliabilities yielded by the separate orders of administration, using the z method described in the foregoing.

Each analogue of the general design thus yielded a set of four correlation matrices, which are presented in this report as indicated in the reference table below. Within each set, the matrices appear in the same order as described above.

Table 4-15. Reference Table for Correlation Matrices

Volume	II	III	IV
Grade	4	5	6
Test	Table Numbers		
Vocabulary	655-658	667-670	679-682
Reading Comprehension	659-662	671-674	683-686
Total Reading	663-666	675-678	687-690

CHAPTER 5

RESULTS

1. Restandardization Study

The Restandardization Study was designed to yield both individual norms and school mean norms for each of the seven reading tests at grades 4, 5, and 6. Test administrations were conducted between April 10 and April 28, 1972, involving a national sample of approximately 200,000 children. The resulting norms are "Spring Norms," and are most appropriate for interpreting the reading performance of children tested at that time.

1.1 Restandardization Sample

The norming design called for the administration of the Metropolitan Reading Test (MAT) in 933 schools with pupils in grades 4, 5, and/or 6. Of these 933 schools, 913 were selected as the primary sample from the school universe data source, and 20 were selected from a special supplementary sample. (For details of the selection procedure, refer to Sections 5 and 7, respectively, of Chapter 3.) Table 5-1 contains the numbers and percentages of primary and substitute schools, by major stratum, that ultimately contributed usable data to the norming study.

Of the 913 schools comprising the primary sample, three schools declined the invitation, and time did not permit the invitation of a substitute. Four schools that had agreed to participate did not administer the test. Of the 20 supplementary schools invited, 7 declined the invitation and no substitutes were available for these schools. One school that had accepted did not administer the test.

TABLE 5-1. ANCHOR TEST STUDY: Restandardization Sample

Numbers and Percentages of Schools* by Major Stratum

Major Stratum	Primary Schools Number	Percentage	Substitute Schools Number	Percentage	Total Number	Percentage
1	17	100.0			17	100
2	16	100.0			16	100
3	14	87.5	2	12.5	16	100
4	10	100.0			10	100
5	12	85.7	2	14.3	14	100
6	9	100.0			9	100
7	9	81.8	2	18.2	11	100
8	24	96.0	1	4.0	25	100
9	12	85.7	2	14.3	14	100
10	11	91.7	1	8.3	12	100
11	37	84.1	7	15.9	44	100
12	7	87.5	1	12.5	8	100
13	27	96.4	1	3.6	28	100
14	10	100.0			10	100
15	10	90.9	1	9.1	11	100
16	11	91.7	1	8.3	12	100
17	7	77.8	2	22.2	9	100
18	10	100.0			10	100
19	10	90.9	1	9.1	11	100
20	8	80.0	2	20.0	10	100
21	16	88.9	2	11.1	18	100
22	12	75.0	4	25.0	16	100
23	14	100.0			14	100
24	40	95.2	2	4.8	42	100
25	10	100.0			10	100
26	19	82.6	4	17.4	23	100
27	32	100.0			32	100
28	36	94.7	2	5.3	38	100
29	113	89.7	13	10.3	126	100
30	13	81.3	3	18.7	16	100
31	9	75.0	3	25.0	12	100
32	42	91.3	4	8.7	46	100
33	14	77.8	4	22.2	18	100
34	12	85.7	2	14.3	14	100
35	10	76.9	3	23.1	13	100
36	11	78.6	3	21.4	14	100
37	12	100.0			12	100
38	16	94.1	1	5.9	17	100
39	11	91.7	1	8.3	12	100
40	10	100.0			10	100
41	14	100.0			14	100
42	13	92.9	1	7.1	14	100
43	39	97.5	1	2.5	40	100
44	10	100.0			10	100
45	10	100.0			10	100
46	10	100.0			10	100
47	7	87.5	1	12.5	8	100
48	12	100.0			12	100
TOTAL	838	91.3	80	8.7	918	100

* These are the schools which provided usable test data.

Tables 5-2, 5-3, and 5-4 contain the numbers and percentages of primary and substitute schools, by state, that participated in the norming study. Table 5-5 shows the numbers of children that participated for the three grades combined. These counts are given in a race x IQ bivariate table with the number of usable, incomplete, invalid, and absentee answer sheets shown for each cell. It is noteworthy that 95% of the children enrolled in the 918 schools provided usable data. While the percentage of usable test data increases with successively higher IQ ranges, the differences among these percentages are small and of no consequence to the norming results. Similarly, the slightly greater percentages of unusable data for the American Indian, Black, and Spanish-surnamed children, relative to the White and Oriental children, are too small to have had a significant effect upon the results.

1.2 Individual and School Mean Norms

The results of the Restandardization Study are shown in Volume I, which contains six sets of tables, two each for grades 4, 5, and 6. The first set of tables at each grade consists of the individual score norms. The second set consists of the school mean norms. Each set of tables is preceded by a list (accessed by a red-lettered tab) describing the order of the tables in that set. The within-set table sequence presents the MAT norms, followed by the norms for the other six tests derived by each of the four procedure/method combinations used in the Equating Study.

Both the individual and school mean norms for the six non-MAT tests were derived by four different procedure/method combinations. However, only those norms resulting from the Average of Procedures 1 and 2, equi-percentile method, are evaluated in terms of their comparability to the

TABLE 5-2. ANCHOR TEST STUDY: RESTANDARDIZATION SAMPLE

Numbers and Percentages of Public Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL	19	86.4	3	13.6	22	100
AK	2	100.0			2	100
AZ	9	90.0	1	10.0	10	100
AR	4	100.0			4	100
CA	65	89.0	8	11.0	73	100
CO	9	100.0			9	100
CT	9	90.0	1	10.0	10	100
DE	3	100.0			3	100
DC	3	75.0	1	25.0	4	100
FL	17	89.5	2	10.5	19	100
GA	12	85.7	2	14.3	14	100
HI	2	100.0			2	100
ID	4	80.0	1	20.0	5	100
IL	31	77.5	9	22.5	40	100
IN	20	83.3	4	16.7	24	100
IA	11	100.0			11	100
KS	7	87.5	1	12.5	8	100
KY	9	100.0			9	100
LA	14	87.5	2	12.5	16	100
ME	5	100.0			5	100
MD	18	94.7	1	5.3	19	100
MA	20	95.2	1	4.8	21	100
MI	39	88.6	5	11.4	44	100
MN	15	88.2	2	11.8	17	100
MS	11	91.7	1	8.3	12	100
MO	16	88.9	2	11.1	18	100
MT	6	100.0			6	100
NE	7	100.0			7	100
NV	2	100.0			2	100
NH	1	100.0			1	100
NJ	29	96.7	1	3.3	30	100
NM	6	100.0			6	100
NY	49	92.5	4	7.5	53	100
NC	26	92.9	2	7.1	28	100
ND	2	100.0			2	100
OH	44	97.8	1	2.2	45	100
OK	4	100.0			4	100
OR	12	92.3	1	7.7	13	100
PA	26	81.3	6	18.7	32	100
RI	5	100.0			5	100
SC	6	100.0			6	100
SD	8	88.9	1	11.1	9	100
TN	21	95.5	1	4.5	22	100
TX	40	87.0	6	13.0	46	100
UT	7	100.0			7	100
VT	3	100.0			3	100
VA	17	85.0	3	15.0	20	100
WA	10	76.9	3	23.1	13	100
WV	3	100.0			3	100
WI	27	100.0			27	100
WY	1	100.0			1	100
TOTAL	736	90.6	76	9.4	812	100

* These are the schools which provided usable test data.

TABLE 5-3. ANCHOR TEST STUDY: RESTANDARDIZATION SAMPLE

Numbers and Percentages of Catholic Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL						
AK						
AZ						
AR						
CA	4	100.0			4	100
CO	1	100.0			1	100
CT	1	100.0			1	100
DE						
DC						
FL	1	100.0			1	100
GA						
HI						
ID						
IL	12	100.0			12	100
IN						
IA						
KS						
KY						
LA	3	75.0	1	25.0	4	100
ME						
MD	3	100.0			3	100
MA	5	100.0			5	100
MI	4	100.0			4	100
MN	6	100.0			6	100
MS	2	100.0			2	100
MO						
MT	1	100.0			1	100
NE						
NV	1	100.0			1	100
NH	1	100.0			1	100
NJ	5	100.0			5	100
NM						
NY	12	92.3	1	7.7	13	100
NC						
ND						
OH	9	90.0	1	10.0		100
OK						
OR						
PA	6	100.0			6	100
RI	1	100.0			1	100
SC						
SD	1	100.0			1	100
TN	2	100.0			2	100
TX	1	100.0			1	100
UT						
VT	1	100.0			1	100
VA						
WA	2	100.0			2	100
WV	1	100.0			1	100
WI	3	100.0			3	100
WY						
TOTAL	89	96.7	3	3.3	92	100

* These are the schools which provided usable test data.

TABLE 5-4. ANCHOR TEST STUDY: RESTANDARDIZATION SAMPLE

Numbers and Percentages of Private Non Catholic Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL						
AK						
AZ						
AR						
CA	2	100.0			2	100
CO						
CT						
DE						
DC						
FL						
GA						
HI						
ID						
IL	2	100.0			2	100
IN	1	100.0			1	100
IA						
KS						
KY						
LA	1	100.0			1	100
ME						
MD						
MA	1	100.0			1	100
MI						
MN	1	100.0			1	100
MS						
MO						
MT						
NE	1	100.0			1	100
NV						
NH						
NJ						
NM						
NY	1	100.0			1	100
NC						
ND						
OH						
OK						
OR	1	100.0			1	100
PA						
RI						
SC						
SD						
TN						
TX						
UT						
VT						
VA						
WA	1	100.0			1	100
WV						
WI	2	100.0			2	100
WY						
TOTAL	14	100.0			14	100

* These are the schools which provided usable test data.

TABLE 5-5. ANCHOR TEST STUDY: RESTANDARDIZATION SAMPLE

Number of Pupils by Racial Groups and IQ Groups with Usable or Absentee, Incomplete, or Invalid Answer Sheets

IQ	Type of Answer Sheet	American Indian	Black	Oriental	Spanish Surname	White or Other	No Information	Total Number	Total Percentage*
		RACIAL GROUPS							
Below 75	Usable	47	2,354	27	409	1,725	12	4,574	90.
	Absentee	4	234	3	45	113		399	
	Incomplete		19		7	11		37	
	Invalid		16			5		25	
75 - 89	Usable	175	6,648	59	1,805	12,641	75	21,403	94.1
	Absentee	14	425	2	140	655	9	1 245	
	Incomplete	1	27		9	25		62	
	Invalid		16			19		35	
90 - 110	Usable	334	7,856	458	2,741	58,629	147	70,165	96.0
	Absentee	22	371	18	150	2,223	10	2,794	
	Incomplete		20		4	67		91	
	Invalid		8		2	27		37	
111 - 125	Usable	127	1,052	278	565	33,948	78	36,048	96.7
	Absentee	7	26	4	21	1,142	3	1,203	
	Incomplete		1		1	24		26	
	Invalid		1			6		7	
Above 125	Usable	47	150	113	78	9,502	28	9,918	96.9
	Absentee		5	1	3	295	2	306	
	Incomplete		1			3		4	
	Invalid					4		4	
No Information	Usable	478	12,186	370	4,230	32,625	752	50,641	93.1
	Absentee	50	733	23	339	1,698	658	3,501	
	Incomplete	2	69	1	34	58	8	172	
	Invalid		27		4	27	1	59	
TOTAL	Usable	1,208	30,246	1,305	9,828	149,070	1,092	192,749	95.1
	Absentee	97	1,794	51	698	6,126	682	9,448	
	Incomplete	3	137	1	55	188	8	392	
	Invalid		68		6	92	1	167	
PERCENTAGE*		92.4	93.8	96.2	92.8	95.9	61.2	95.1	

* The percentage of usable answer sheets for each marginal cell (i.e., racial group or IQ level).

publishers' norms since only the equivalency tables derived from this method will be of general use.

Before presenting a detailed comparison of Anchor Test Study and publishers' norms, however, it should be pointed out that not all test publishers provide percentile ranks for each raw score point, or for each of the three reading scores. Therefore comparisons of ITBS and SAT total reading norms and STEP II vocabulary and reading comprehension norms were not possible. In those few instances in which the test publishers did not provide percentile ranks for each individual score point, a linear interpolation was applied to adjacent percentile ranks to determine score point percentile ranks. This procedure was followed to facilitate a single, uniform comparison between the Anchor Test Study norms and the test publishers' norms.

The Anchor Test Study norms for each of the six non-MAT tests were derived by applying the results of the Equating Study to the Anchor Test Study MAT norms. Because of the small equating error estimates, the consistency of equating results across procedures and methods of equating, and the generally high degree of comparability among the seven reading tests, the norms for each of these tests may be considered as valid as those of the MAT itself.

1.3 Comparisons Between Anchor Test Study and Test Publishers' Norms

Tables 2149 through 2169 of Volume XXX show the Anchor Test Study norms and the corresponding publisher's norms, as well as the difference between corresponding pairs of percentile ranks. These data may be summarized as follows:

(a) The CAT publisher's norms differ from the Anchor Test Study norms by no more than 3 percentile rank points. Generally, the publisher's percentile ranks are higher for scores below the median, and about the same or lower for scores above the median.

(b) The CTBS publisher's norms differ from the Anchor Test Study norms by no more than 4 percentile rank points, except at grade 6. At grades 4 and 5, the publisher's percentile ranks are generally higher for scores below the median, and about the same or lower for scores above the median. At grade 6, the publisher's percentile ranks differ from those of the Anchor Test Study by no more than 6 percentile rank points, and they are generally lower for scores both above and below the median.

This means that, in general, children who score below the median on the CAT (at all three grades) and on the CTBS (at grades 4 and 5) tend to look slightly better on the publisher's norms than they would on the Anchor Test Study norms, while the opposite would tend to be generally true for children who score above the median. For the CTBS at grade 6, however, the publisher's norms generally yield a slightly lower national standing for most children than would the Anchor Test Study norms.

(c) The ITBS publisher's norms differ from the Anchor Test Study norms by no more than 7 percentile rank points. The publisher's percentile ranks are generally lower throughout most of the score range.

(d) The MAT publisher's norms differ from the Anchor Test Study norms by no more than 3 percentile rank points, except at grade 4 where the largest difference is 8 percentile ranks. The publisher's percentile ranks are generally lower, but not consistently so

throughout the score range, except at grade 4. At grade 4, the publisher's norms were obtained through the use of scorable test booklets while the Anchor Test Study norms were obtained through the use of a separate answer sheet.

(e) The STEP Series II publisher's norms differ from the Anchor Test Study norms by no more than 5 percentile rank points. The publisher's percentile ranks are generally lower throughout the score range, except at grade 6 in the third quartile range, where the publisher's percentile ranks are slightly higher.

(f) The SRA publisher's norms differ from the Anchor Test Study norms by no more than 5 percentile rank points. The publisher's percentile ranks are generally lower throughout most of the score range.

(g) The SAT publisher's norms differ from the Anchor Test Study norms by as much as 18 percentile rank points. The publisher's percentile ranks are consistently lower, except at the extremes of the score range, and in the top quartile of grade 4. In these cases, the publisher's percentile ranks are about the same or slightly higher than those of the Anchor Test Study norms.

For these five tests, the publisher's norms generally yield lower national standing than would the Anchor Test Study norms for most children at the three grades. As indicated in the detailed comparisons above, these differences range from negligible to rather substantial.

Despite the fact that test publishers, in norming their own reading tests, depend on both smaller and perhaps somewhat less representative samples than that of the Anchor Test Study, these comparisons show that the differences between the normative data independently obtained by each publisher and the

REVISED

normative data obtained in the Anchor Test Study are, with one exception, essentially minimal. The one exception is the Stanford Achievement Tests; the publisher's norms would consistently yield significantly lower percentile ranks for SAT scores throughout most of the range than would use of the Anchor Test Study norms.

2. Equating Study

The primary objective of the Equating Study was to equate each of seven reading tests to each of the other six tests at grades 4, 5, and 6. The subsequent development of equivalency tables would make it possible to translate a child's score on any of the seven tests into an equivalent score on any of the other tests. The study was also designed to investigate the consistency among the equating results derived from various equating procedures and methods as well as those obtained for different random and nonrandom subgroups of the total equating sample. The test administrations for the Equating Study were also conducted during the period of April 10 to April 28, 1972, involving a national sample of approximately 140,000 children.

2.1 Equating Sample

The equating design, which is schematically represented in Table 4-4, required 448 school-units,* 16 school-units for each of 28 distinct pairs of tests. While 447 school-units actually tested, two school-units

---

* For a definition of school-unit, see footnote, Section 2, Chapter 4.

conducted their administration in such a way that their data were judged to be invalid, and consequently were dropped from the study. Tables 5-6 through 5-9 show the numbers and percentages of schools that comprise the 445 school-units ultimately included in the Anchor Test Equating Study. Table 5-6 shows the numbers and percentages of schools from the primary and substitute samples by each of 27 major strata. Tables 5-7, 5-8, and 5-9 show the composition of the primary and substitute samples by state, for public, Catholic, and private schools. Only 13% of the schools that participated in the study were substitute schools. Table 5-10 shows the numbers of children that participated in the Equating Study for the three grades combined. The counts are given in a race x IQ bivariate table, with the number of usable, incomplete, invalid, and absentee answer sheets shown for each cell. It is noteworthy that 94% of the children enrolled in the 736 schools provided usable test data. While the percentage of usable test data increases with successively higher IQ ranges, the differences are small, and of no consequence for the equating results. Similarly, the observed differences in percentage of usable data for racial subgroups are too small to have had a significant effect upon the results.

2.2 Comparability of the Tests

While the equating of parallel test forms has been common practice among test publishers for many years, the Anchor Test Equating Study represents the first attempt to equate different tests, not designed as parallel forms. Because these tests vary in length and degree of difficulty, have different ceilings and floors, and were constructed from various philosophical bases, it is appropriate to consider the question of whether such equating is warranted. To support the inference that

TABLE 5-6. ANCHOR TEST STUDY: Equating Sample

Numbers and Percentages of Schools* by Major Stratum

Major Stratum	Primary Schools Number	Primary Schools Percentage	Substitute Schools Number	Substitute Schools Percentage	Total Number	Total Percentage
1	15	88.2	2	11.8	17	100
2	45	97.8	1	2.2	46	100
3	29	85.3	5	14.7	34	100
4	22	100.0			22	100
5	37	80.4	9	19.6	46	100
6	15	62.5	9	37.5	24	100
7	11	100.0			11	100
8	18	90.0	2	10.0	20	100
9	14	82.4	3	17.6	17	100
10	17	100.0			17	100
11	8	80.0	2	20.0	10	100
12	25	89.3	3	10.7	28	100
13	20	83.3	4	16.7	24	100
14	35	81.4	8	18.6	43	100
15	31	86.1	5	13.9	36	100
16	25	92.6	2	7.4	27	100
17	37	97.4	1	2.6	38	100
18	13	92.9	1	7.1	14	100
19	7	77.8	2	22.2	9	100
20	15	78.9	4	21.1	19	100
21	30	88.2	4	11.8	34	100
22	32	97.0	1	3.0	33	100
23	32	80.0	8	20.0	40	100
24	16	80.0	4	20.0	20	100
25	40	93.0	3	7.0	43	100
26	33	73.3	12	26.7	45	100
27	17	89.5	2	10.5	19	100
TOTAL	639	86.8	97	13.2	736	100

* These are the schools which provided usable test data.

TABLE 5-7. ANCHOR TEST STUDY: EQUATING SAMPLE

Numbers and Percentages of Public Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL	17	81.0	4	19.0	21	100
AK						
AZ	3	75.0	1	25.0	4	100
AR	3	60.0	2	40.0	5	100
CA	45	83.3	9	16.7	54	100
CO	3	100.0			3	100
CT	5	41.7	7	58.3	12	100
DE	3	100.0			3	100
DC	2	66.7	1	33.3	3	100
FI	18	94.7	1	5.3	19	100
GA	8	75.0	4	25.0	12	100
HI			1	100.0	1	100
ID	5	100.0			5	100
IL	31	96.9	1	3.1	32	100
IN	9	100.0			9	100
IA	5	100.0			5	100
KS	10	100.0			10	100
KY	7	87.5	1	12.5	8	100
LA	15	88.2	2	11.8	17	100
ME	2	100.0			2	100
MD	4	100.0			4	100
MA	15	88.2	2	11.8	17	100
MI	14	70.0	6	30.0	20	100
MN	7	77.8	2	22.2	9	100
MS	11	100.0			11	100
MO	18	94.7	1	5.3	19	100
MT	11	100.0			11	100
NE	8	88.9	1	11.1	9	100
NV						
NH	2	66.7	1	33.3	3	100
NJ	7	58.3	5	41.7	12	100
NM	7	77.8	2	22.2	9	100
NY	45	86.5	7	13.5	52	100
NC	22	91.7	2	8.3	24	100
ND	2	100.0			2	100
OH	27	96.4	1	3.6	28	100
OK	6	100.0			6	100
OR	8	88.9	1	11.1	9	100
PA	17	70.8	7	29.2	24	100
RI			2	100.0	2	100
SC	9	75.0	3	25.0	12	100
SD	7	100.0			7	100
TN	11	91.7	1	8.3	12	100
TX	31	100.0			31	100
UT	2	66.7	1	33.3	3	100
VT						
VA	17	85.0	3	15.0	20	100
WA	15	93.8	1	6.2	16	100
WV	10	90.9	1	9.1	11	100
WI	10	90.0	1	9.1	11	100
WY	2	100.0			2	100
TOTAL	536	86.3	85	13.7	621	100

* These are the schools which provided usable test data.

TABLE 5-8. ANCHOR TEST STUDY: EQUATING SAMPLE

Numbers and Percentages of Catholic Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL	2	100.0			2	100
AK						
AZ	4	100.0			4	100
AR						
CA	8	100.0			8	100
CO						
CT	3	100.0			3	100
DE						
DC						
FL	1	100.0			1	100
GA						
HI						
ID						
IL	6	100.0			6	100
IN	4	100.0			4	100
IA	4	100.0			4	100
KS						
KY	2	100.0			2	100
LA						
ME						
MD						
MA	4	100.0			4	100
MI	7	100.0			7	100
MN	3	100.0			3	100
MS						
MO	11	91.7	1	8.3	12	100
MT						
NE						
NV						
NH						
NJ	4	80.0	1	20.0	5	100
NM						
NY	4	80.0	1	20.0	5	100
NC						
ND						
OH	2	100.0			2	100
OK						
OR	2	100.0			2	100
PA	10	100.0			10	100
RI						
SC						
SD						
TN						
TX						
UT						
VT						
VA	2	100.0			2	100
WA						
WV	4	100.0			4	100
WI	1	100.0			1	100
WY						
TOTAL	88	96.7	3	3.3	91	100

* These are the schools which provided usable test data.

TABLE 5-9. ANCHOR TEST STUDY: EQUATING SAMPLE

Numbers and Percentages of Private Non Catholic Schools* by States

State	Primary Schools		Substitute Schools		Total	
	N	%	N	%	N	%
AL						
AK						
AZ						
AR						
CA	4	80.0	1	20.0	5	100
CO						
CT	1	33.3	2	66.7	3	100
DE						
DC						
FL	1	100.0			1	100
GA						
HI						
ID						
IL	1	100.0			1	100
IN						
IA						
KS	1	50.0	1	50.0	2	100
KY						
LA						
ME						
MD	2	100.0			2	100
MA			1	100.0	1	100
MI			1	100.0	1	100
MN						
MS						
MO	2	50.0	2	50.0	4	100
MT						
NE						
NV						
NH						
NJ						
NM						
NY	1	100.0			1	100
NC						
ND						
OH						
OK						
OR						
PA	1	100.0			1	100
RI						
SC						
SD						
TN						
TX						
UT						
VT						
VA						
WA	1	50.0	1	50.0	2	100
WV						
WI						
WY						
TOTAL	15	62.5	9	37.5	24	100

* These are the schools which provided usable test data.

TABLE 5-10. ANCHOR TEST STUDY: EQUATING SAMPLE

Number of Pupils by Racial Groups and IQ Groups with Usable or Absentee, Incomplete, or Invalid Answer Sheets

IQ	Type of Answer Sheet	RACIAL GROUPS: American Indian	Black	Oriental	Spanish Surname	White or Other	No Information	Total Number	Total Percentage*
Below 75	Usable	66	2,431	3	310	1,113	24	3,947	90.2
	Absentee	9	186	1	27	71		294	
	Incomplete	2	59		13	38	2	114	
	Invalid		11		1	8	1	21	
75 - 89	Usable	167	5,799	3	1,203	8,416	51	15,679	92.5
	Absentee	15	311	1	81	417	7	832	
	Incomplete	4	161		46	173	1	385	
	Invalid		16		2	34		54	
90 - 110	Usable	272	6,156	347	1,952	38,524	117	47,368	94.9
	Absentee	[illegible]	224	4	85	1,318	12	1,658	
	Incomplete	5	153	5	61	572	4	800	
	Invalid		12			57	2	71	
111 - 125	Usable	69	792	221	477	23,295	42	24,896	96.0
	Absentee	2	19	4	13	654	6	698	
	Incomplete		14	1	8	287	4	314	
	Invalid		1	1		11		13	
Above 125	Usable	15	131	62	71	6,877	17	7,173	96.0
	Absentee	1	7	3	3	192	1	207	
	Incomplete		7	2	3	77		89	
	Invalid					3	1	4	
No Information	Usable	227	7,895	235	3,184	23,506	745	35,792	91.7
	Absentee	42	526	10	278	965	287	2,108	
	Incomplete	12	346	5	134	489	62	1,048	
	Invalid		25		6	35	5	71	
TOTAL	Usable	816	23,204	907	7,197	101,731	1,000	135,855	93.9
	Absentee	84	1,273	23	487	3,617	313	5,797	
	Incomplete	23	740	13	265	1,636	73	2,750	
	Invalid		67	1	9	148	9	234	
PERCENTAGE*		88.4	91.8	96.1	90.4	95.0	71.7	93.9	

* The percentage of usable answer sheets for each marginal cell (i.e., racial group or IQ level).

achievement tests of this type do measure essentially the same intellectual functions and reading skills, and thus that their equating is justified, they should generally have intercorrelations, corrected for unreliability of the tests, of .95 or higher, i.e., their true score correlations should approach 1.0.

On the basis of this arbitrary but widely accepted criterion, the equating of these seven tests is generally supported by the summary correlation data in Table 5-11 and the last column of Tables 5-12, 5-13, and 5-14. Of the 189 corrected correlations, 159 are .95 or higher and 10 are .94. Thus, only 20 corrected correlations (about 10%) would be considered slightly less than the desired value, and these lower correlations, moreover, tend to be distributed across tests and grades and not consistently associated with any test-pair. In addition, it may be noted in Tables 5-12, 5-13, and 5-14 that the observed correlations for each test cluster about the alternate forms reliability of the test itself.

2.3 Equating Results for the Total Sample

The results of the Anchor Test Equating Study, for the total sample, appear in Volumes II through X. Volumes II, III, and IV, for grades 4, 5, and 6, respectively, are identically organized, each containing five sets of tables. The first three sets in each volume are the equating tables for vocabulary, reading comprehension, and total reading scores, in that order. The fourth set consists of the estimated equating error tables, and the fifth set, the tables of correlations. Each of the five sets is preceded by a list (accessed by a red-lettered tab) describing the order of appearance of the tables in that set. For the first three sets, the within-set table sequence is the same, presenting the equating tables produced by

TABLE 5-11. ANCHOR TEST EQUATING STUDY

Distributions of Correlations Corrected for Attenuation

Grade 4

Range of Corrected Intercorrelations	Vocabulary	Comprehension	Total Reading	Across Tests
.98 - 1.00	13	5	14	32
.95 - .97	5	10	7	22
.92 - .94	3	5	0	8
.89 - .91	0	1	0	1

Grade 5

Range of Corrected Intercorrelations	Vocabulary	Comprehension	Total Reading	Across Tests
.98 - 1.00	10	7	11	28
.95 - .97	7	7	8	22
.92 - .94	3	7	2	12
.89 - .91	1	0	0	1

Grade 6

Range of Corrected Intercorrelations	Vocabulary	Comprehension	Total Reading	Across Tests
.98 - 1.00	16	15	15	46
.95 - .97	2	3	4	9
.92 - .94	2	2	2	6
.89 - .91	1	1	0	2

Grades 4, 5, & 6 Combined

Range of Corrected Intercorrelations	Vocabulary	Comprehension	Total Reading	Across Tests
.98 - 1.00	39	27	40	106
.95 - .97	14	20	19	53
.92 - .94	8	14	4	26
.89 - .91	2	2	0	4

TABLE 5-12. ANCHOR TEST EQUATING STUDY

Reliabilities and Ranges of Averaged Intercorrelations Among Tests: Grade 4

Test	Alternate-Forms Reliability	Range of Observed Intercorrelations *	Range of Corrected Intercorrelations *
Vocabulary			
CAT	.88	.82 - .87	.97 - 1.00
CTBS	.87	.84 - .88	.96 - 1.00
ITBS	.87	.82 - .89	.93 - 1.00
MAT	.88	.81 - .88	.93 - 1.00
STEP II	.80	.81 - .85	.97 - 1.00
SRA	88	.82 - .89	.93 - 1.00
SAT	.82	.81 - .87	.97 - 1.00
Comprehension			
CAT	.83	.76 - .82	.93 - .98
CTBS	.86	.78 - .86	.93 - 1.00
ITBS	.88	.78 - .86	.92 - 1.00
MAT	.86	.76 - .86	.91 - 1.00
STEP II	.80	.76 - .82	.91 - 1.00
SRA	.84	.79 - .86	.94 - 1.00
SAT	.84	.80 - .83	.96 - .98
Total Reading			
CAT	.91	.88 - .90	.96 - 1.00
CTBS	.92	.88 - .92	.96 - 1.00
ITBS	.93	.88 - .93	.95 - 1.00
MAT	.92	.86 - .92	.95 - 1.00
STEP II	.89	.86 - .91	.95 - 1.00
SRA	.91	.87 - .93	.95 - 1.00
SAT	.89	.88 - .90	.98 - 1.00

* The intercorrelation ranges for each score indicate the minimum and maximum intercorrelations for the given test with each of the other tests.

TABLE 5-13. ANCHOR TEST EQUATING STUDY

Reliabilities and Ranges of Averaged Intercorrelations Among Tests: Grade 5

Test	Alternate-Forms Reliability	Range of Observed Intercorrelations *	Range of Corrected Intercorrelations *
Vocabulary			
CAT	.89	.81 - .89	.94 - 1.00
CTBS	.89	.80 - .89	.92 - 1.00
ITBS	.87	.82 - .90	.95 - 1.00
MAT	.88	.82 - .90	.96 - 1.00
STEP II	.83	.75 - .86	.89 - 1.00
SRA	.90	.82 - .89	.93 - 1.00
SAT	.85	.75 - .84	.89 - .96
Comprehension			
CAT	.86	.78 - .83	.92 - .97
CTBS	.87	.80 - .85	.92 - .98
ITBS	.87	.82 - .89	.94 - 1.00
MAT	.86	.82 - .89	.95 - 1.00
STEP II	.84	.78 - .85	.92 - 1.00
SRA	.86	.81 - .87	.95 - 1.00
SAT	.86	.80 - .86	.92 - .99
Total Reading			
CAT	.92	.88 - .91	.95 - .99
CTBS	.93	.86 - .92	.93 - .99
ITBS	.92	.88 - .94	.95 - 1.00
MAT	.93	.89 - .94	.96 - 1.00
STEP II	.90	.85 - .92	.93 - 1.00
SRA	.92	.88 - .93	.96 - 1.00
SAT	.91	.85 - .90	.93 - .98

* The intercorrelation ranges for each score indicate the minimum and maximum intercorrelations for the given test with each of the other tests.

TABLE 5-14. ANCHOR TEST EQUATING STUDY

Reliabilities and Ranges of Averaged Intercorrelations Among Tests: Grade 6

Test	Alternate-Forms Reliability	Range of Observed Intercorrelations *	Range of Corrected Intercorrelations*
Vocabulary			
CAT	.85	.79 - .89	.95 - 1.00
CTBS	.87	.77 - .89	.91 - 1.00
ITBS	.87	79 - .88	.92 - 1.00
MAT	.89	.84 - .89	.98 - 1.00
STEP II	.83	.77 - .84	.91 - .98
SRA	.89	.84 - .89	.97 - 1.00
SAT	.88	.80 - .89	.93 - 1.00
Comprehension			
CAT	.82	.76 - .83	.91 - .99
CTBS	.85	.79 - .86	.95 - 1.00
ITBS	.85	.77 - .86	.92 - 1.00
MAT	.86	.77 - .86	.91 - 1.00
STEP II	.84	.76 - .84	.92 - 1.00
SRA	.85	.82 - .86	.98 - 1.00
SAT	.87	.83 - .86	.98 - 1.00
Total Reading			
CAT	.90	.85 - .91	.94 - 1.00
CTBS	.91	.86 - .93	.95 - 1.00
ITBS	.92	.85 - .91	.93 - 1.00
MAT	.93	.88 - .92	.97 - 1.00
STEP II	.91	.85 - .90	.93 - .98
SRA	.92	.90 - .93	.98 - 1.00
SAT	.92	.88 - .92	.96 - 1.00

* The intercorrelation ranges for each score indicate the minimum and maximum intercorrelations for the given test with each of the other tests.

the eight procedure/method combinations. For the fourth and fifth sets, tables are sequenced by vocabulary, reading comprehension, and total reading tests.

Volumes V through X contain the equating line graphs for the total sample, comparing the various procedures and methods by which they were generated. Each volume is prefaced by a list of the abbreviations, symbolisms, and cross-references for the graphs contained therein. Volumes V, VII, and IX, for grades 4, 5, and 6, respectively, contain the graphs comparing the four equating procedures, for each method of equating, linear and equipercentile. These three volumes are identically organized, each consisting of six sets of graphs, two sets each for vocabulary, reading comprehension, and total reading, in that order. Each pair of sets, preceded by a list (accessed by a red-lettered tab) describing the order of appearance of the graphs, compares equating lines for the Average of Procedures 1 and 2 vs Procedure 3 and Procedure 4 vs Procedure 2. Within each set, procedure comparisons for the linear method (odd-numbered graphs) and for the equipercentile method (even-numbered graphs) appear alternately.

Volumes VI, VIII, and X, for grades 4, 5, and 6, contain the graphs comparing the linear and equipercentile equating methods for each of the four procedures. These three volumes are identically organized, each consisting of three sets of graphs, one each for vocabulary, reading comprehension, and total reading, each preceded by a list (accessed by a red-lettered tab) describing the order of appearance of the graphs within the set. Each set is divided into seven subsets, one for each test, which compare linear and equipercentile equating lines for each of Procedures 1, 2, 3, and the Average of Procedures 1 and 2, respectively, within each test-pair.

2.4 Evaluation of Equating Procedures and Methods

The equating of each of the seven tests was carried out by three procedures, each representing a different way of organizing and utilizing the available data to which the linear and equipercentile methods of equating were applied. Procedures 1 and 2, because each used both first and second test data, interchanging their roles analogously, were expected to yield essentially the same equating results, which were to be subsequently averaged. Procedure 3, by simply combining the data for each subgroup that took a given test without regard to order of administration, was intended to produce equating results in a manner similar to that which test publishers have generally used in conducting equating studies, to provide a basis for comparing the results of averaging Procedures 1 and 2.

The equating results obtained by Procedures 1 and 2 were indeed consistent, and thus the averaging of the corresponding equating tables was justified. Furthermore, there was general agreement between the equating results of the Average of Procedures 1 and 2 and those of Procedure 3. Procedure 3 utilized largely nonoverlapping samples which were considerably smaller than those for the Average of Procedures 1 and 2, which was explicitly designed to minimize sampling variance by providing a common data base over the entire equating sample. Thus, Procedure 3 would be expected to be more sensitive to random sampling fluctuations, an expectation supported by the results of the equating error analyses. Accordingly, the equating results obtained by the Average of Procedures 1 and 2 are judged to be superior to those obtained by Procedure 3.

With regard to the results obtained from the linear and equipercentile methods of equating, those obtained by the equipercentile method are judged to be superior. The moments (the mean, variance, skewness, and kurtosis)

of the score distributions for the seven tests differed substantially. Thus, a linear transformation based only on the mean and variance of each distribution was generally inadequate, often yielding equivalent scores which fell outside the corresponding raw score range of the tests.

2.5 Equating Error

The tables of estimated error of equating for the total sample are presented in Volumes II, III, and IV for grades 4, 5, and 6, respectively. The equating error estimates were obtained for the vocabulary, reading comprehension, and total reading scores for each of the six non-MAT tests when equated to the corresponding scores only of the MAT, by each of the eight different procedure/method combinations. Thus, no estimated equating error is provided for the MAT or for the other six tests when equated to each other. Because the same score distributions upon which these estimates were obtained would have been used in estimating the equating error for each test when equated to every other test, the obtained estimates are inferred to be generally comparable to those that would have been obtained if the error analysis had been carried out for all equatings of each test. Furthermore, while the equating error estimates generated by the eight different procedure/method combinations were of theoretical importance and provided an additional basis for evaluating all the equating results, only those obtained for the Average of Procedures 1 and 2, equipercentile method, are of practical importance, since the equivalency tables derived in this particular manner are the ones that will be generally used. Tables 5-15, 5-16, and 5-17 contain a summarized version of the estimated equating error for the Average of Procedures 1 and 2, equipercentile method. The individual score point

TABLE 5-15. ANCHOR TEST EQUATING STUDY: GRADE 4

Estimates of Averaged* Error in Equating the Reading Tests to the MAT (70 ED.) Elementary, Form F

Average of Procedures 1 and 2, Equipercentile

Raw Score Range	CAT(70 ED.) Level 3 Form A Average Error	CTBS(68 ED.) Level 2 Form Q Average Error	ITBS(71 ED.) Level 10 Form 5 Average Error	STEP II(69 ED.) Level 4 Form A Average Error	SRA(71 ED.) Blue Form E Average Error	SAT(64 ED.) Intermediate Form W Average Error
Vocabulary						
41 - 45					.1	
36 - 40	.1	.1			.1	.1
31 - 35	.1	.1	.1		.1	.1
26 - 30	.1	.1	.1	.1	.1	.1
21 - 25	.2	.3	.2	.1	.2	.2
16 - 20	.5	.6	.4	.2	.3	.4
11 - 15	1.0	1.1	.9	.6	.8	.8
6 - 10	.7	.8	.7	.7	.7	1.0
1 - 5	.6	.4	.6	.6	.5	.7
Reading Comprehension						
66 - 70			.1			
61 - 65			.1			
56 - 60			.2			.1
51 - 55			.2			.2
46 - 50			.2		.1	.2
41 - 45	.1	.1	.2		.1	.2
36 - 40	.2	.2	.2		.1	.2
31 - 35	.2	.2	.3		.1	.2
26 - 30	.2	.3	.3	.1	.2	.4
21 - 25	.3	.3	.3	.2	.3	.5
16 - 20	.4	.3	.4	.2	.3	.5
11 - 15	.5	.3	.4	.3	.4	.5
6 - 10	.5	.4	.6	.5	.4	.5
1 - 5	.5	.5	.8	.5	.5	.6
Total Reading						
101 - 110			.2			
91 - 100			.2			.3
81 - 90	.2	.2	.2		.2	.2
71 - 80	.2	.2	.2		.1	.2
61 - 70	.2	.2	.3		.2	.2
51 - 60	.2	.4	.5	.2	.3	.4
41 - 50	.4	.4	.5	.1	.4	.6
31 - 40	.7	.5	.7	.3	.5	.7
21 - 30	1.1	1.0	1.1	.5	1.0	1.4
11 - 20	1.1	1.0	1.1	.9	1.0	1.4
1 - 10	2.1	2.5	2.5	.9	.7	1.6

* For vocabulary and reading comprehension, error estimates were averaged over intervals of five raw score points and smoothed. For the total reading test, estimates were averaged over intervals of ten and then smoothed.

TABLE 5-16. ANCHOR TEST EQUATING STUDY: GRADE 5

Estimates of Averaged* Error in Equating the Reading Tests to the MAT (70 ED.) Intermediate, Form F

Average of Procedures 1 and 2, Equipercentile

Raw Score Range	CAT(70 ED.) Level 3 Form A Average Error	CTBS'68 ED.) Level 2 Form Q Average Error	ITBS(71 ED.) Level 10 Form 5 Average Error	STEP-II(69 ED.) Level 4 Form A Average Error	SRA(71 ED.) Blue Form E Average Error	SAT(64 ED.) Intermediate Form W Average Error
Vocabulary						
46 - 50						.1
41 - 45			.1			.1
36 - 40	.1	.1	.2		.1	.1
31 - 35	.2	.2	.2		.2	.2
26 - 30	.3	.3	.2	.2	.3	.2
21 - 25	.3	.3	.2	.2	.3	.3
16 - 20	.3	.3	.3	.3	.3	.3
11 - 15	.3	.3	.3	.3	.4	
6 - 10	.4	.5	.3	.5	.6	.6
1 - 5	.4	.9	.5	.6	.7	.7
Reading Comprehension						
71 - 75			.1			
66 - 70			.1			
61 - 65			.1			.1
56 - 60			.1			.2
51 - 55			.2			.2
46 - 50			.2		.1	.2
41 - 45	.1	.1	.2		.2	.2
36 - 40	.2	.2	.2		.2	.2
31 - 35	.2	.2	.2		.2	.2
26 - 30	.2	.2	.2	.1	.2	.2
21 - 25	.2	.2	.2	.2	.2	.2
16 - 20	.2	.2	.2	.2	.2	.2
11 - 15	.2	.2	.3	.2	.2	.2
6 - 10	.2	.3	.6	.2	.2	.4
1 - 5	.5	.6	.6	.4	.4	.4
Total Reading						
111 - 120			.3			.4
101 - 110			.3			.4
91 - 100			.3			.2
81 - 90	.3	.3	.3		.3	.3
71 - 80	.3	.2	.3		.3	.3
61 - 70	.3	.4	.3		.3	.3
51 - 60	.4	.4	.3	.3	.3	.3
41 - 50	.5	.5	.4	.3	.4	.4
31 - 40	.5	.5	.4	.3	.4	.5
21 - 30	.4	.5	.4	.4	.4	.5
11 - 20	.4	1.2	.8	.7	.9	.8
1 - 10	.4	2.1	2.0	.9	2.1	1.2

* For vocabulary and reading comprehension, error estimates were averaged over intervals of five raw score points and smoothed. For the total reading test, estimates were averaged over intervals of ten and then smoothed.

TABLE 5-17. ANCHOR TEST EQUATING STUDY: GRADE 6

Estimates of Averaged* Error in Equating the Reading Tests to the MAT (70 Ed.) Intermediate, Form F

Average of Procedures 1 and 2, Equipercentile

Raw Score Range	CAT(70 ED.) Level 3 Form A Average Error	CTBS(68 ED.) Level 2 Form Q Average Error	ITBS(71 ED.) Level 10 Form 5 Average Error	STEP-II(69 ED.) Level 4 Form A Average Error	SRA(71 ED.) Blue Form E Average Error	SAT(6[illegible] ED.) Intermediate Form W Average Error
Vocabulary						
46 - 50			.1			.1
41 - 45			.1		.[illegible]	.1
36 - 40	.1	.1	.[illegible]		.1	.2
31 - 35	.1	.1	.1		.2	.2
26 - 30	.1	.1	.2	.1	.2	.3
21 - 25	.2	.2	.3	.1	.3	.3
16 - 20	.3	.3	.4	.3	.3	.4
11 - 15	.3	.5	.4	.5	.3	.4
6 - 10	.4	.5	.6	.6	.6	.5
1 - 5	.6	.9	1.1	.9	.8	.7
Reading Comprehension						
76 - 80			.1			
71 - 75			.1			
66 - 70			.1			
61 - 65			.1			.1
56 - 60			.1			.1
51 - 55			.2			.1
46 - 50			.2		.1	.1
41 - 45	.1	.1	.2		.1	.2
36 - 40	.1	.2	.2		.1	.2
31 - 35	.1	.2	.2		.2	.3
26 - 30	.2	.2	.2	.1	.2	.3
21 - 25	.2	.2	.3	.2	.2	.3
16 - 20	.2	.3	.3	.2	.2	.3
11 - 15	.4	.3	.3	.3	.3	.3
6 - 10	.4	.3	.4	.3	.3	.3
1 - 5	.8	.5	.5	.3	.7	.5
Total Reading						
121 - 130			.3			
111 - 120			.3			.2
101 - 110			.2			.2
91 - 100			.1			.1
81 - 90	.2	.2	.2		.2	.1
71 - 80	.2	.1	.2		.2	.2
61 - 70	.2	.2	.2		.2	.4
51 - 60	.2	.2	.3	.2	.3	.4
41 - 50	.2	.3	.4	.3	.3	.5
31 - 40	.4	.4	.7	.4	.4	.5
21 - 30	.6	.6	.7	.6	.5	.6
11 - 20	.6	.6	1.2	.8	.8	.8
1 - 10	1.9	2.8	1.8	1.7	1.2	1.2

*For vocabulary and reading comprehension, error estimates were averaged over intervals of five raw score points and smoothed. For the total reading test, estimates were averaged over intervals of ten and then smoothed.

estimates were averaged* and smoothed to reduce the slight fluctuations at adjacent raw score points and to facilitate the use of such estimates. On the basis of these estimates, it is concluded that the equating error associated with any score outside what is generally referred to as the "chance score range" is but a fraction (generally a small fraction) of a raw score point. Therefore, for individual score interpretations, the error of equating would be considered to be a trivial factor, relative to the error of measurement. For group data, such as school means, districts means, etc., the equating error generally would tend to average a half raw score point or less, assuming that most test scores fall above the mean chance level.

2.6 Subgroup Equating Results

The equating results for the subgroup analyses are contained in Volumes XI through XXVII. Each of Volumes XI through XXI presents an entire set of equating tables for one of eleven special subgroups. These volumes are identically organized; in each, the tables appear successively by grade, and within each grade, by vocabulary, reading comprehension, and total reading tests, resulting in nine sets of equating tables per volume. Each set contains the Procedure 3, linear method tables, ordered by test, followed by Procedure 3, equipercentile method tables, similarly ordered.

Volumes XXII through XXVII contain the graphs comparing the subgroup equating lines for each of four breakdowns of the total sample. Each

---

* For vocabulary and reading comprehension, the individual raw score point estimates were averaged over a 5 point interval, and those for the total reading were averaged over a 10 point interval.

volume is prefaced by a list of the abbreviations, symbolisms, and cross-references for the graphs contained therein. Graphs generated by Procedure 3, linear method appear in Volumes XXII, XXIV, and XXVI for grades 4, 5, and 6, respectively, while those produced by Procedure 3, equipercentile method are contained in Volumes XXIII, XXV, and XXVII, for the three grades. These volumes are identically organized each presenting the comparisons for the four breakdowns in the following order: sex, SES, IQ, and race. Within each type of breakdown, the graphs are sequenced by vocabulary, reading comprehension, and total reading scores.

Evaluation of Subgroup Equating Results: The primary means of evaluating the results for the various special subgroups was by examination of the graphs comparing subgroup equating lines. Since the standard error of equating based on the total sample is not applicable to the subgroup analyses (and statistical tests are thus not available), differences between subgroups of a given type were appraised chiefly on the basis of consistency of deviation in a given direction from the equating line for the total sample.

For the sex breakdown, the equating lines are virtually identical in almost all instances. Hence, the equating tables for the total sample may be considered equally applicable to both sexes.

For the analyses by socioeconomic levels, some differences were noted. In equating the CTBS to each of the other tests, there is a consistent tendency for the high SES group equating curves to be lower than those for the medium and low SES groups and for the total group, implying that the high SES children score consistently higher on the CTBS than on any of the other tests. This observation persists across vocabulary, reading comprehension, and total reading tests at all grades. On the other

hand, the low SES children, at all three grades, score consistently somewhat lower on both subtests and total test for the SRA than on the other tests. These observed differences may be due to the fact that SES subgroups were formed in terms of school characteristics, not pupil characteristics as was true for all other subgroups. Thus the reliability for individual pupil assignment to the three SES subgroups might be lower than would be expected if individual pupil data had been available.

The equating results for the IQ subgroups generally parallel those of the total equating sample. No systematic differences between the equating lines of the IQ subgroups and the total equating group were found for all tests or for a given test across the three grades. For most tests at grade 4, the low IQ equating lines for the vocabulary scores in the top 1/3 of the score range show the greatest tendency to deviate from those obtained for the total equating group. However, generally less than 5% of the low IQ children have scores in this upper range. The differences observed at grade 4 tend to diminish at successively higher grades where larger percentages of low IQ children have scores in the top 1/3 of the score range, presumably because of both improved reading ability and/or diminishing difficulty level of the test forms used at the higher grades. Therefore, it would seem reasonable to infer that lack of data in the top 1/3 of the score range might be the major relevant factor associated with observed differences between the equating lines of the low IQ subgroup and the total equating group. Lack of data would also seem to be the most appropriate explanation for the observed differences at the lower end of the score range for the high IQ subgroup at grade 6.

The racial subgroup results tend in many instances, to deviate markedly from those for the total equating sample, for some tests, and

in some portion of the score range. In particular, for the Spanish-surnamed subgroup, the equating lines for the ITBS to each of the other tests are consistently and substantially above those for the total group through the top 2/3 of the score range, and those for the SRA are consistently above those for the total group through the top 1/2 or 1/3 of the range. Both trends are apparent in the vocabulary, reading comprehension, and total reading results and are maintained over all three grades. This results from the fact that, in comparison with the total equating sample, there are relatively fewer Spanish-surnamed children scoring in the upper portion of the score range on both of these tests than on the other five tests.

For the Black subgroup, the deviations are neither so pronounced as are those for the Spanish-surnamed subgroup nor are they generally consistent over grades for subtest or total test scores. For example, in the results for the vocabulary subtest at the 4th grade, there are notable deviations for the CTBS, ITBS, and SAT in the uppermost portion of the score range, but these differences disappear at the 5th and 6th grades. For reading comprehension, the most substantial differences at grade 4 occur for the equating of the ITBS and SAT to each of the other tests through most of the score range. At the 5th grade, the deviations are restricted to the top half of the range for ITBS only, while at the 6th grade, only at the very top of the SAT range are there any notable discrepancies. For total reading results at the 4th grade, the equating lines for Blacks are below those for the total sample in the top half of the CTBS and SAT score range, while in the 5th and 6th grades, differences are most substantial at the uppermost portion of the ITBS range.

It is particularly important to note that the study was not explicitly designed to yield stable equating relationships for the minority subgroup children as those for the total equating group. Therefore, a substantial portion of the observed discrepancies is likely to be due to small samples, particularly in the higher score range where the data are quite sparse. In the case of the Black subgroup, generally less than 10% of the children score in the upper half of the score range, while for the Spanish-surnamed subgroup, generally less than 10% of the children obtain scores in the upper 1/3 of the score range. Furthermore, the average subgroup size for the Spanish-surnamed children is quite small, about 1/4 the size of the Black subgroup. Thus, the number of Spanish-surnamed children who score at any one point of the score range is indeed small and subject to relatively large random sampling fluctuations. Nevertheless the observed trends are such that they warrant consideration and may not be able to be dismissed on the basis of insufficient data and random sample fluctuations. The use of the racial subgroups equivalency score data, however, is not generally recommended because the observed differences may be partially attributable to rather large random sampling variations.

CHAPTER 6

RECOMMENDATIONS

1. Selection of the Anchor Test

The central role played by the test which serves as the "anchor" for the Equating Study makes it appropriate to consider the criteria upon which the selection of such a test should be based. Because the distribution of the test used as the base test in the frequency estimation procedure greatly affects the estimated marginals generated thereby, and because the equipercentile method of equating forces the effective score range of all other tests to be comparable to that of the anchor test to which they are each equated, it is particularly important for the anchor test score distribution to be symmetric over its range with "ideal" floor and ceiling.

While the MAT was generally satisfactory in this regard, the score distribution at grade 4 was somewhat negatively skewed and the test appears to have less than an ideal ceiling. Thus, it is here proposed that the selection of the test to serve as anchor test be based upon an ad hoc examination of the distributions obtained and/or a thorough examination of the various publishers' norms and norming procedures.

2. Changes in Equating Procedures and Methods

Given the results of procedure and method comparisons made possible by the present study, it is recommended that future analyses using this design be carried out only by the equipercentile method and only for Procedures 1 and 2, which may then be averaged for the final equating tables.

The linear method proved to be generally unsatisfactory because the distributions of tests designed to measure achievement at different multiple grade combinations are not similar in shape or effective range. Linear transformations of curvilinearly related tests cannot, in principle, be satisfactory, particularly at the extremes of the score range.

While it is true that the Procedure 3 results obtained here were generally comparable to those of the Average of Procedures 1 and 2, the estimated error of equating associated with Procedure 3 was generally larger than that for the Average of Procedures 1 and 2, and thus the results of Procedure 3 are less reliable.

3. Generalizability of Equating Results to Special Subgroups

The subgroup equating analyses were intended to permit a limited evaluation of the generalizability of the total sample equating results to these subgroups. As would be expected from previous experience, the results for boys and girls were essentially the same. The generalizability would also appear to extend to the IQ subgroups, where the results were comparable to those of the total group except in the score ranges where the data were sparse.

For the racial subgroups, however, and to a lesser extent for the SES subgroups, the results are somewhat inconclusive (see Section 2.6 of Chapter 5). Although the sampling plan used in the present study could not be expected to have yielded larger samples of Black, Spanish-surnamed, and high SES children, the possibility of further analyses of the present subgroup data to draw more conclusive inferences about the applicability of the total sample results should be investigated.

4. Operational and Administrative Recommendations

(a) At least six additional working months will be important to the execution of another study. Three of these months should be provided for the conduct of the invitational phase, and three more months for the materials design phase. The latter would also provide more time for the required USOE and OMB forms clearance.

(b) Since one of the most difficult and time-consuming parts of the study was contact with potential participants, serious consideration should be given to including more than one subject area and/or additional grade levels. Once initial contact has been established with the school, the addition of another hour of class time and/or the inclusion of additional grades requires relatively little effort and cost.

(c) It is suggested that future studies combine the district and school invitation phases. All school invitation packets should accompany the district level invitation. The District Superintendent, after approving participation and appointing a District Coordinator, can immediately assure the distribution of the school invitations and accompanying materials. By again enclosing the Pre-Test Information Form in the school invitation packets, several mailings and follow-ups can be eliminated without affecting the participation rate.

(d) The status reporting of invitations, receipts, score reports, etc. was originally specified as being required on a daily or weekly basis. It is our recommendation that future studies call for such reporting on a bi-/tri-weekly basis.

(e) The number of ancillary forms specified in future projects should be drastically reduced as accomplished in this study. All information required for the conduct of the Anchor Test Study was obtained from the Pre-Test Information Form, Class Identification Form, School Coordinator's Report Form, and the answer sheet(s). It was possible to eliminate (with USOE approval) a number of specified forms without loss of either data or convenience.

(f) The contractor should plan to conduct all follow-ups by telephone. Personal contact will help to assure a high participation rate.

(g) Serious consideration should be given to utilization of the administrative/operational format as developed for this Anchor Test Study (described in this Final Report). Use of the current contractor's expertise should also play a major role in the conduct of such projects.

(h) It would be helpful to include on the answer sheet an area for the contractor's staff to enter a disposition code for all "irregular" cases reported by the test administrator or school coordinator. This would greatly facilitate the later determination of the inclusion of such cases in the score reporting and/or analytic phases of the study.

(i) The "For Test Administrator Use Only" area of the answer sheet should be expanded to cover severe emotional (not just physical or mental) disorders (question 1) and to solicit an opinion regarding the effect of such a disorder upon the validity of the test score (through an additional question).

(j) Directions should be provided to the test administrator to cover the common irregularities which will arise: how to handle cases

where the child answers in the booklet instead of the answer sheet or loses his place on the answer sheet; what to do if a child becomes ill or disruptive to the rest of the class; etc.

APPENDICES

A. The Determination of Approximately Optimum Probabilities of Selection for Size Class of School

B. The Determination of the Set of Measures of Size

C. The Formation of Major Strata

D. Approximation of District-Selection Probabilities

E. Survey Estimates and Corresponding Variance Estimates for the Restandardization Study

F. Classification Procedure for "Weighting Classes" and Definitions of Eligibility and Usable Cases

G. Tukey-Cureton Smoothing Procedure

H. Notation

J. Frequency Estimation Procedure

K. Equipercentile Equating Procedure

L. Computational Formulas

APPENDIX A

THE DETERMINATION OF APPROXIMATELY OPTIMUM PROBABILITIES OF SELECTION BY SIZE CLASS OF SCHOOL

A.1 The Choice of Optimum Sampling Fractions by Size of School for the Restandardization Study

This section of Appendix A examines for the Restandardization Study the effect of an approximately optimum allocation of the sample by size class of school, and thus provides the basis for determining the measures of size and the probabilities of selection that were used in the selection of the sample.

The tentative data used in this memorandum on the size distribution of schools were derived from the 1968-69 school universe tape. This study was done before similar results were available from the 1970-71 school universe tape and from other sources that were used in actual sample selection. The results presented in this Appendix would be altered in the details, but not in their general character, as a result of the revisions that would be made with the revised data.

Since all students in the specified grades within the selected schools were to be administered the Anchor Test, both the contribution to the variance of an average test score or percentile, and the cost of including a school in the sample would be substantially influenced by the size of the eligible enrollment in a school. In this situation, strata based on size of schools might have been used, along with other stratification criteria.

Actually, after ascertaining the sampling fractions for optimum allocation of the sample to size strata, we used these results to determine measures of size and to assign varying probabilities of selection to schools proportionate to their measures of size, and thus to achieve the optimum sampling fractions without actually using size as a criterion for stratification. We were then able to utilize other criteria more fully for stratification, and still reduce the effects of varying school sizes, and also achieve approximately optimum allocation of the sample by size of school. The size of a school is the number of students in it in grades 4, 5, and 6, and this can be approximated on the basis of the information on the 1970-71 school universe tape. There will be no bias, but a larger variance, if the approximation to the number of pupils eligible to take the test in April, 1972, is not a reasonably good one for some of the schools.

A.1.1 The Theory*

The percentile to be estimated is T_P, the test score such that the desired proportion P of pupils has a test score less than T_P, with

$$P = \frac{X}{Y}$$

where

X is the total number of pupils in the specified grade in the universe

Y is the total number of pupils in the specified grade.

---

* See Hansen, Hurwitz and Madow, Sample Survey Methods and Theory, Vol. I, pp. 179-237 and 448-449.

We can estimate T_p from the sample by finding t_p, the score such that

$P = x'/y'$ is the desired proportion observed in the sample,

where

y' is the sample estimate of the total eligible pupils in the specified grade, and

x' is the sample estimate of the number of these pupils with scores less than t_p.

For a stratified sample with the schools within a size stratum selected with equal probability, and with the test given to all eligible students in each selected school, the sample estimates, x' and y', are here assumed to be of the form

$$x' = \sum_{h}^{L} \frac{N_h}{n_h} \sum_{i}^{n_h} x_{hi} \text{ , and}$$

$$y' = \sum_{h}^{L} \frac{N_h}{n_h} \sum_{i}^{n_h} y_{hi} \text{ ,}$$

where

N_h is the total number of schools in stratum h,

n_h is the number of schools in the sample from stratum h,

x_{hi} is the number of eligible students in the i^{th} sampled school in stratum h with scores less than t_p,

y_{hi} is the number of eligible students in school hi, and

L is the number of strata.

In practice, the population frequency distribution of scores is estimated from the sample by estimating frequencies for scores or intervals of scores, using the same type of formula as that above for x'. Then, percentiles are easily estimated from the estimated cumulative frequency distribution by interpolation.

Confidence limits can be computed for T_p, based on the standard deviation σ_p of the sample estimate, p, of the proportion of students who score less than T_p. This proportion p is a ratio-type estimate with variance*

$$\sigma_p^2 \doteq \frac{1}{\bar{Y}^2} \sum_h^L \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{S_{Ph}^2}{n_h},$$

where

$$S_{Ph}^2 = \sum_i^{N_h} (Z_{hi} - \bar{Z}_h)^2 / (N_h - 1),$$

$$Z_{hi} = X_{hi} - PY_{hi},$$

$$\bar{Z}_h = \sum_i^{N_h} Z_{hi} / N_h,$$

X_{hi} = the number of eligible students in school i of stratum h with a test grade below the percentile T_p (the population parameter), and

Y_{hi} = the total number of eligible students in school i of stratum h.

---

* See Ibid., p. 190, equation (4.5), and the remark on p. 193.

The upper and lower confidence bounds, t_{P_U} and t_{P_L}, for a 68-percent confidence interval for the percentile T_P will be approximately the sample estimates of the P_U^{th} and P_L^{th} percentiles, where*

$$P_U = P + \sigma_p \text{ , and}$$

$$P_L = P - \sigma_p \text{ .}$$

An approximate cost function for a stratified sample of the type proposed for the standardization study, with all eligible students in a sampled school taking the test, is

$$C = C_0 + \sum_h^L \bar{C}_h n_h \text{ ,}$$

where

C_0 represents costs that do not vary with the number or sizes of the schools in the sample, and

$\bar{C}_h$ is the average cost associated with each school included in the sample in size class h.

---

* See Ibid., pp. 448-449.

The maximum precision of results will be achieved for a given number of schools in the sample of schools if the sample is allocated to the size strata such that the sample from the h^{th} size class is*

$$n_h = n \frac{N_h S_{Ph} / \sqrt{\bar{C}_h}}{\sum_h^L N_h S_{Ph} / \sqrt{\bar{C}_h}} .$$

Here we make the assumption that, for any percentile T_P,

$$S_{Ph} \doteq k_P S_h ,$$

where expressions for k_P and S_h will be derived later. Under this assumption, the expression for the optimum n_h becomes

$$n_h = n \frac{N_h S_h / \sqrt{\bar{C}_h}}{\sum_h^L N_h S_h / \sqrt{\bar{C}_h}} .$$

The approximation $S_{Ph} \doteq k_P S_h$ implies that the S_{Ph} vary in the same relative manner between the size strata, for any percentile. When this holds, the optimum allocation of the sample will be the same for the estimation of each percentile. The basis for this assumption will be indicated later.

In practice we do not know the S_h in advance of obtaining the sample survey results, but can often develop reasonably good approximations to them, as demonstrated later. Also we can, at best, only approximate the cost function. Nevertheless, even very rough approximations can be highly useful in arriving at an approximately optimum allocation of the sample.

---

* See Ibid, p. 221.

There is no bias in the sample if the approximations are in error, although the variance is somewhat larger than it would be with the values known. However, unless the estimated gains from optimum allocation to strata are substantial and the approximations used are in the right direction, it may be desirable to use uniform fractions, since the effort to accomplish an optimum allocation could result in a loss over proportionate sampling, instead of a gain. We shall see later that the estimated gains were modest, but to achieve these modest gains and for other reasons it was recommended and decided to use varying probabilities by size of school that approximated the effect of optimum allocation of the sample to size strata.

We shall now obtain approximate values for S_h and $\bar{C}_h$. Actually, as can be seen from the formula for the optimum n_h, it is not necessary to obtain the absolute values, but only numbers proportionate to them, in order to allocate the sample in an optimum manner.

A.1.2 Approximate Values for the Cost Function

In the Restandardization Study a tentative approximate value for $\bar{C}_h$ is given by $\bar{C}_h = 100 + .85\ \bar{E}_h$ dollars,

where

$\bar{E}_h$ is the average number of tests to be administered to eligible pupils per school in size class h.

(Note that $\bar{E}_h$ is the average per school of all eligible pupils in grades 4, 5, and 6 combined, and is different from $\bar{Y}_h$ which is the average number per

$$P_h = \frac{\sum_i^{N_h} X_{hi}}{\sum_i^{N_h} Y_{hi}} = P$$

for each of the size classes. Under this assumption, $\bar{Z}_h = 0$, and therefore the S^2_{Ph} (defined earlier) becomes

$$S^2_{Ph} \doteq \sum_i Z^2_{hi}/(N_h - 1)$$

$$= \sum_i (X_{hi} - PY_{hi})^2/(N_h - 1)$$

$$= \sum_i Y^2_{hi}(P_{hi} - P)^2/(N_h - 1) ,$$

where

$$P_{hi} = X_{hi}/Y_{hi}.$$

Moreover, for schools of size Y_{hi}, the expected value of $(P_{hi} - P)^2$ is

$$E(P_{hi} - P)^2 = \frac{PQ}{Y_{hi}}\left\{1 + \rho_{Phi}(Y_{hi} - 1)\right\} ,$$

where

$$\rho_{Phi} = \frac{E(X_{Phij} - P)(X_{Phij'} - P)}{E(X_{Phij} - P)^2}$$

X_{Phij} = 1 if the j^{th} pupil in the specified grade of school i, size-class h, has a score below T_p.

= 0 otherwise,

$X_{Phij'}$ is the corresponding value for any <u>other</u> pupil in the same grade of the same school, and

$P_{hi} = \sum_j X_{hij}/Y_{hi} = X_{hi}/Y_{hi}$ is the average value of X_{hij} within school i, i.e., the proportion of students in school i with scores less than T_p.

school of pupils in a specified grade for which a percentile is to be estimated. The $\bar{Y}_h$ appears in the variance approximations and the $\bar{E}_h$ in the cost function.)

The values assumed in this cost function were obtained through consultation with ETS. The $100 is the fixed cost for each school in the sample, independent of the number of tests administered. It is based on a fixed fee of $40 to be paid to a coordinator for each sampled school, plus other costs speculated to be of the order of $60, including the costs of obtaining cooperation of a school in the survey, the costs of providing instructions and controls, and other costs that vary directly with the number of schools included in the sample. The .85 $\bar{E}_h$ in the cost function is the allowance for the number of tests to be administered in the school. The differences in the sample design will not be seriously affected if this fixed cost per school included in the sample is off by as much as, say, 25 percent.

A.1.3 Approximate Values for the Variances, S^2_{Ph}

We understand from the analyses of data assembled in the Coleman report* and from other sources that the proportion of students with scores less than T_p does not vary widely between size classes of schools. If we make this assumption, then

---

* George Mayeske et al., A Study of Our Nation's Schools (a U.S. Office of Education working paper analyzing the data collected for the Coleman report).

In the above equation, ρ_{Phi} is intraclass correlation of the X_{Phi}, among pupils within schools, for schools of size Y_{hi}.

Another way of stating ρ_{Phi} in terms of the total variance between pupils and the variance between school means is

$$\rho_{Phi} = \frac{\sigma_b^2 - \sigma^2/Y_{hi}}{\sigma^2} \quad \frac{Y_{hi}}{Y_{hi} - 1}$$

where

Y_{hi} is the size of the grade in the hi^{th} school,

σ^2 is the total variance between pupils, and

σ_b^2 is the variance between school means.

The subtractive term in the numerator will be negligible when Y_{hi} is reasonably large.

It is now assumed that the intraclass correlation, ρ_{Phi}, takes on approximately the same value, ρ_P, for each size of school. Actually, experience shows us for many different variables that on the average the intraclass correlation for clusters of different sizes ordinarily decreases slowly as size of cluster increases.

With the above assumption, and substituting Y_{hi} for $Y_{hi} - 1$, the expression for $E(P_{hi} - P)^2$ is then approximately

$$E(P_{hi} - P)^2 = \frac{PQ}{Y_{hi}} (1 + \rho_P Y_{hi}).$$

If the above expression is substituted for $(P_{hi} - P)^2$ and if N_h is substituted for $N_h - 1$, the equation for S^2_{Ph} becomes

$$S^2_{Ph} = PQ \sum_{i}^{N_h} Y_{hi}(1 + \rho_P Y_{hi})/N_h$$

$$= PQ \{\bar{Y}_h + \rho_P \bar{Y}^2_h (1 + V^2_h)\} ,$$

where

V^2_h = the relvariance among schools within size stratum h of enrollment in the particular grade for which a percentile is computed.

If we now make, at least for the present, the further assumption that ρ_P is the same for all percentiles (i.e., $\rho_P = \rho$), we have

$$S^2_{Ph} = PQ \left[\bar{Y}_h + \rho \bar{Y}^2_h (1 + V^2_h)\right] .$$

This assumption seems plausible, and there is a limited amount of data available in an unpublished memorandum in the Office of Education files to support the above assumption.*

We now have the evidence for the assumption made earlier where we assumed that we could approximate S_{Ph} by $k_P S_h$. In the result we have just obtained, the value of k_P is $\sqrt{PQ}$, and

$$S_h = \sqrt{\bar{Y}_h + \rho \bar{Y}^2_h (1 + V^2_h)} .$$

---

* For SAT scores for sixth graders in a recent study in Palo Alto, California, the Office of Education examined the variance between students of the Y_{ij} variable (defined above) for the 10th, 50th, and 90th percentiles. For each of these percentiles, the variance between schools was about 1/3 of the total variance between students. Hence, the intraclass correlation ρ_P was approximately equal to 1/3 for each of these percentiles.

A.1.4 Computations

We shall now apply these results to approximate data obtained in the tabulation of the 1968-1969 school universe tape referred to earlier.

Table A-1 shows, by size class of combined enrollment in grades 4, 5, and 6, estimates of the number of schools, total combined enrollment in grades 4, 5, and 6, average enrollment in the combined classes, and a rough estimate of the enrollment per grade. It also shows an approximate value for $1 + V_h^2$. The V_h^2 values were not computed from the data but were roughly approximated from the width of the class intervals and the means for the size distributions of enrollment given in Table A-1. Nevertheless, they are likely to be reasonably good approximations.

Table A-2 shows the optimum allocation of the sample and compares it with proportionate allocation for a sample of 1,200 schools (this is the size of sample initially called for by the contract for the study, and based on the assumption that schools would be sampled with equal probability). We have carried this computation through for two values of ρ: $\rho = .15$ and $\rho = .20$. The use of values of .15 and .20 for ρ are based on information in a memorandum from John Bianchini dated August 4, for STEP Series II Reading. Those results are also consistent with results given by Angoff* in which he reported a variance between school mean enrollments equal to approximately 25 percent of the variance between pupils. The ratio of between-school variance to total variance of 25 percent would yield an intraclass correlation of .25.

---

* Robert L. Thorndike, ed., Educational Measurement (2nd Edition), American Council on Education (1971), chapter 15.

Table A-1. Preliminary size distribution of public schools based on enrollment in grades 4, 5, and 6*

School Size Class (combined enrollment in grades 4, 5, and 6)	Schools		Total Enrollment in Grades 4, 5, and 6	Average Enrollment per School, Grades 4, 5, and 6	Average Enrollment per Grade (rough)	$1 + V_h^2$ where V_h^2 equals Relvariance of Enrollment
	Number N_h	Percent of Total	E_h	$\bar{E}_h$	$\bar{Y}_h$	
Less than 50	11,537	18	247	21	7	1.36
50-99	10,584	16	798	75	25	1.03
100-199	18,667	29	2,782	149	50	1.03
200-499	22,347	35	6,505	291	97	1.06
500-999	1,293	2	775	617	206	1.10
1,000 +	29	.04	41			
Totals	64,457	100	11,147			

* These results were obtained by using 1968-1969 school tapes, for schools containing grades 4, 5, or 6, and imputing grade enrollments by assuming that the enrollment in grades 4, 5, and 6 contained in a school is $\frac{g}{t} e$,

where

g is the number of grades 4, 5, and 6 in the school,

t is the number of grades 1 through 12 in the school, and

e is the enrollment in these grades.

The $\bar{Y}_h$ was approximated simply as $\bar{E}_h/3$.

Table A-2. Comparison of allocations of the sample, variances and costs

School Size Class	Number of Schools N_h	Number of Schools in Sample, n_h, for Alternative Allocations		
		Optimum Allocation		Proportionate Allocation
		ρ = .20	ρ = .15	
Less than 50	11,537	49	55	215
50-99	10,584	114	119	197
100-199	18,667	348	345	348
200-499	22,347	630	622	416
500 +	1,322	60	60	25
Totals	64,457	1,200	1,200	1,200
Variance (x k) of Estimated Norm for Sample of 1,200 Schools*		.731	.564	ρ = .20 .930 ρ = .15 .715
Variable Cost ($1,000)		398	395	325
Average Number of Pupils in Sample per Grade per School in Sample**				

* The values in the table are computed as

$$\sum_h^{N_h} \frac{N_h^2}{N^2} \cdot \frac{N - n_h}{N_h} \cdot \frac{S_h^2}{n_h} .$$

Therefore, since $S_{Ph}^2 = PQS_h^2$, the variance, σ_p^2, for a particular percentile T_p, can be obtained approximately by multiplying the tabled value by $\frac{PQ}{Y^2}$.

** Only the ratios of these numbers were obtained from this study. They are adjusted to an assumed average of 67 for all schools.

The results summarized in A Study of Our Nation's Schools, cited earlier, show approximately 35 percent of the total variance accounted for by schools. This variance is presented as an upper-limit estimate and, again, is for an achievement measure and not just reading. They report essentially the same proportion of variance accounted for by schools for each of the grades studied.

For these computations, we have assumed that stratification of schools would have the effect of reducing the proportion of total variance accounted for by between-school variance to the alternate assumed levels of .20 and .15. From the evidence in A Study of Our Nation's Schools, it seems clear that reductions of this order of magnitude and, in fact, considerably greater reductions could be achieved if we had adequate variables for use in stratification or estimation that are related to average economic and social characteristics of the families served by the schools. A discussion of the stratification criteria actually used appears elsewhere. Here we simply assume that we can obtain some substantial gains from stratification. Consequently, we have chosen the indicated assumed values for the intraclass correlation.

A.1.5 Results

From Table A-2, it is seen that if $\rho = .20$, the variance (x k) for a sample of 1,200 schools is reduced from .930 for proportionate allocation of the sample to .731 for optimum allocation, but at an increase in variable costs of from $297,000 to $360,000. The gain is relatively more than the increase in cost, but not strikingly so. Thus, if the proportionate sample

were increased in size from 1,200 schools to 1,450, the cost of the proportionate sample would be the same as a sample of 1,200 schools with optimum allocation. The variance (x k) for a proportionate sample of this size would be .766. Thus we see that for a given size of sample there is a substantial reduction in variance, but for equal costs the gain is a moderate one of approximately 5 percent.

A consequence of using optimum allocation instead of proportionate allocation is a substantial decrease in the number of smaller schools in the sample, and a corresponding increase in the larger schools. This was regarded as an advantage, over and above the net cost gains. The result is an increase from an average of 67* persons per school to be tested to 91 students per school to be tested.** This, of course, is the cause of the increase in cost for the optimum allocation of the sample.

In practice, we will not know the standard deviation exactly, and the gain from optimum allocation may be somewhat less than we have estimated. However, in the nature of this problem, there are reasons to assume that we should be able to approximate the optimum reasonably well.

As seen in Table A-2, the results for $\rho = .15$ are essentially similar to those for $\rho = .20$, but with considerably smaller variances as a result of the reduced correlation.

---

* This figure was not developed from our preliminary tabulations, but is the figure used by OE and ETS in current discussions at the time of this analysis.

** From the data used in this analysis, the estimated average size of school for proportionate sample was 57.7, and for optimum allocation with $\rho = .20$ was 78.3. The estimate of 91 was obtained by computing 67 x 78.3/57.7.

A decision was made, on the basis of these results, to adopt measures of size and probabilities of selection proportionate to these measures of size that would yield sampling fractions by size of school approximately equivalent to the optimum sampling fractions with size of school strata. A decision was also made to reduce the sample size from the initially contemplated 1,200 schools to 940. It was estimated that this size of sample would provide smaller variances at the same cost as compared with the original plan for 1,200 schools selected with equal probability. The reduction in sample size allowed not only for the use of varying and approximately optimum probabilities, but also for some additional costs to introduce stratification variables not already on the OE school universe tapes.

A.2 Variances and Allocation of Sample by School Size Classes for the Equating Study

Two methods are proposed for the equating of tests: the equal percentile method, and the linear method.

A.2.1 Variance of the Equipercentile Method

The equipercentile method for equating tests A and B involves equating the test scores on A and B for any specified percentile rank. This is equivalent, for example, to specifying a test score A, for test A, estimating its percentile rank, and then equating to the score B, on test B, with the same estimated percentile rank.

If both tests A and B are administered to all of the students in the same sample of schools (it is here assumed that both tests are administered to each student without practice effects), then the variance of the percentile B, given a specified percentile A, will be determined approximately by and will be approximately proportionate to

$$\sigma^2_{p} = \sigma^2_{p_A} + \sigma^2_{p_B} - 2\rho_{p_A p_B}\sigma_{p_A}\sigma_{p_B}$$

where

$$\rho_{p_A p_B} = \sigma_{p_A p_B}/\sigma_{p_A}\sigma_{p_B}$$

is the coefficient of correlation of p_A and p_B and

$$\sigma_{p_A p_B} = \frac{1}{\bar{Y}^2}\sum_h^L \frac{N_h^2}{N^2}\,\frac{N_h - n_h}{N_h n_h}\,S_{(Ph)AB}$$

with

$$S_{(Ph)AB} = \sum_i^{N_h} (Z_{Ahi} - \bar{Z}_{Ah})(Z_{Bhi} - \bar{Z}_{Bh})/(N_h - 1)$$

is the covariance of $Z_{Ahi} = X_{Ahi} - PY_{hi}$ and $Z_{Bhi} = X_{Bhi} - PY_{hi}$ within stratum h.

We can rewrite the variance of p in the form

$$\sigma^2_{p} = \frac{1}{\bar{X}^2}\sum_h^L \frac{N_h^2}{N^2}\,\frac{N_h - n_h}{N_h n_h}\,S^2_{PDh}$$

where

$$S^2_{PDh} = S^2_{PAh} + S^2_{PBh} - 2S_{(Ph)AB}$$

is the variance of the difference between Z_{Ahi} and Z_{Bhi}

$$= S^2_{PAh} + S^2_{PBh} - 2\rho_{(Ph)AB}S_{PAh}S_{PBh}\,.$$

The variance of the estimated percentile rank, p_A, associated with the specified test score A is (see Section 1.1 above)

$$\sigma^2_{p_A} = \frac{1}{\bar{Y}^2} \sum_h^L \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h} \frac{S^2_{pAh}}{n_h}$$

where

$$S^2_{pAh} = \sum_i^{N_h} (Z_{hi} - \bar{Z}_h)^2/(N_h - 1)$$

$$Z_{hi} = X_{ri} - PY_{hi},$$

and the notation used is defined in Section 1.1 with, as before, Y_{hi} equal to the number of students in school i within size stratum h.

There is a percentile B (on test B) that corresponds, on the average, in successive samples to the specified percentile A (on test A). The variance of the estimate of this percentile, if estimated from an independent sample, will be approximately proportionate to the variance $\sigma^2_{p_B}$, which is defined in the same manner as $\sigma^2_{p_A}$. If tests are given to all of the students in independent samples of schools, selected within size of school strata, the confidence bounds of the equating scores on test B for a given score on Test A will be determined approximately by the variance

$$\sigma^2_p = \sigma^2_{p_A} + \sigma^2_{p_B}.$$

That is, the confidence bounds of the estimated percentile for B for a selected value of test score A will be determined approximately by the range of scores on test B encompassed by $p \pm t\sigma_p$. Consequently, the standard error of the percentile will be approximately proportionate to σ_p.

If the intraclass correlation of scores among students within schools for test B is approximately the same as for test A, and both are approximately equal to ρ_p, for each stratum as defined and discussed in Section 1, and with the assumptions and results from Section 1, we have

$$s^2_{1h} = 2PQ\{\bar{Y}_h + \rho_p \bar{X}^2_h(1 + V^2_h)\}(1 - \rho_{(Ph)AB}) .$$

If we assume, as we did in Section 1, that $\rho_p = \rho$, that is, the intraclass correlation is about the same for the different percentiles, and if we also assume that roughly $\rho_{(Ph)AB}$ is about the same for the different percentiles and for the different size strata (this assumption seems plausible and a reasonable one for guiding sample design when we have multiple statistics to estimate from the sample), then we have

$$\rho_{AB} = \rho_{(Ph)AB}$$

and

$$S^2_{Dh} = 2PQ(1 - \rho_{AB})\{\bar{Y}_h + \rho\bar{Y}^2_h(1 + V^2_h)\} .$$

It follows, if the above assumptions are reasonable ones, that optimum allocation of the sample to size classes of schools for the Equating Study should be the same (adjusted to different level of sample desires) as for the restandardization sample. We believe the assumptions are plausible ones, at least for equipercentile equating.

A.2.2 Linear Equating

It seems reasonable and we have some tentative results that suggest that a similar design will be relatively efficient for linear equating also.

Moreover, even if this does not hold, we would assume that the equipercentile equating approach should be given priority in terms of sampling efficiency of the estimates. Of course consistent estimates can be made by either method from a probability sample--the only goal here is to minimize the variances for a given cost.

APPENDIX B

THE DETERMINATION OF THE SET OF MEASURES OF SIZE

In order to determine approximate optimum sample sizes for schools in various enrollment classes, an approximate optimum allocation analysis was carried out. This analysis is summarized in Section 3 of the report and described in detail in Appendix A. The resulting approximate optimum sample sizes and corresponding sampling fractions for schools in the enrollment classes are given in the table below.

Table B-1. Approximate optimum sample sizes

Enrollment in Grades 4, 5, and 6	Approximate Number of Schools in Universe * (N_j)	Approximate Optimum Sample Size (n_j)	Approximate Optimum Sampling Fraction (f_j)
Less than 50	11,537	38	.0033
50 - 99	10,584	89	.0084
100 - 199	18,667	273	.0146
200 - 499	22,347	493	.0221
500 or more	1,322	47	.0356
Total	64,457	940	

* The frequencies in this column were obtained from the 1968-1969 school universe tape. In cases where the school enrollment was not given grade by grade, the enrollment in grades 4, 5, and 6 was estimated as:

$$(g/t)e ,$$

where

g = the number of grades 4, 5, and 6 in the school,

t = the number of grades 1 through 12 in the school, and

e = the enrollment in the grades 1 through 12 in the school.

As discussed earlier in detail, a decision was made not to stratify schools by enrollment-size class in order to allow for more extensive use of other variables in defining strata. However, the probability of selecting schools from each of the five enrollment classes was controlled by assigning measures of size to schools based on their enrollment class. That is, measures of size were assigned to the schools in such a way that the expected sampling fraction (or selection probability) for the schools in each enrollment class would be approximately equal to those given in the last column of Table B-1.

The measures of size that were assigned to schools to achieve the purpose described above were derived from the fact that the expected sample size for a particular enrollment class is equal to the total sample size (940 schools) times the proportion of the total of the measures of size for all schools accounted for by those in the particular enrollment class. That is, if N_j represents the total number of schools in the j^{th} enrollment class, n_j represents the sample size for the j^{th} enrollment class, and A_j represents the measure of size assigned to each school in the j^{th} enrollment class, then the following equation holds:

$$E(n_j) = 940\ (N_j A_j / A), \qquad j = 1, 2, 3, 4, 5, \tag{1}$$

where

$$A = \sum_{j}^{5} N_j A_j = \text{the total of the measures of size of all schools in the universe.}$$

If both sides of equation 1 are divided by N_j, the following expression for the selection probability (or expected sampling fraction, f_j) for the j^{th} enrollment class is obtained:

$$E(f_j) = 940\ (A_j/A), \quad j = 1, 2, 3, 4, 5 . \qquad (2)$$

Equation 2 shows that the expected sampling fraction for the schools in a particular enrollment class is proportional to the measure of size assigned to the schools in that enrollment class. Therefore, measures of size proportional to the approximate optimum sampling fractions given in Table B-1 were assigned to the schools in each enrollment class. This was done by multiplying the f_j values in Table B-1 by 1,000 and rounding to whole numbers. In the process of rounding, it was decided to round upward for the smaller size classes and downward for the larger size classes. This decision was based on experience with previous studies, which indicates that larger variances than anticipated often arise for smaller size classes.*

Any set of positive values approximately proportional to the f_j values given in Table B-1 could have been used for the set of measures of size. The set chosen was one that was convenient. In particular, this set contained measures of size with a maximum of two digits and that were large enough so that whole numbers could be used without a large amount of rounding, on a percentage basis, of the measures of size assigned to each group.

---

* See Hansen, Hurwitz, and Madow, Sample Survey Methods and Theory, Vol. I, p. 218.

Table B-2 below shows the set of measures of size used along with the approximate optimum sampling fractions (i.e., the f_j values from Table B-1). From this table, it can be noted for each of the five enrollment classes how the product of 1,000 and the approximate optimum sampling fraction were rounded to the whole number used as the measure of size.

Table B-2. Measures of size

Enrollment in Grades 4, 5, and 6	Approximate Optimum Sampling Fraction (f_j)	Measure of Size Used (A_j)
Less than 50	.0033	4
50 - 99	.0084	9
100 - 199	.0146	16
200 - 499	.0221	22
500 or more	.0356	35

APPENDIX C

THE FORMATION OF MAJOR STRATA

As described in Sections 5.3 and 6.2 of this report, major strata were formed for both the Restandardization Study and the Equating Study by collapsing adjacent cells in several tables. Schools were placed in cells in these based upon measures of income level and percent minority described earlier.

Tables C-1 through C-3 show how the cells were grouped to form 40 major strata in the three tables used for public schools. Aggregate measures of size are given for each cell. Table C-4 indicates the grouping of schools into six major strata for Catholic schools by combining schools in adjacent income classes. The number of each major stratum is also indicated in these tables by the circled number in each cluster of cells. No table is given for non-Catholic private schools since only one major stratum (number 41) was defined for them.

Tables C-5 through C-7 show the formation of the 22 major strata defined among public schools for the Equating Study. Table C-8 shows the definition of four major strata among Catholic schools. Again, the aggregate measure of size for each cell and the number of each final stratum are given in these tables. The single major stratum for non-Catholic private schools (number 6) does not appear in the tables.

The aggregate measure of size for each major stratum and the number of final strata defined in each major stratum were given earlier in Table 2 for all the 47 major strata in the Restandardization Study and in Table 3 for the 27 major strata in the Equating Study.

Table C-1. Aggregate measures of size for the Restandardization Study by cell for public schools in the 50 largest school districts (major stratum groups and numbers are indicated)

Percent Minority	Average Adjusted Gross Income (in thousands)							
	2 - 2.9	3 - 3.9	4 - 4.9	5 - 5.9	NA*	6 - 7.9	8 - 9.9	10+
<5	0	138	1,053	4,927	491	12,791	7,649	915
5 - 9	0	86	467	1,929	120	5,843	2,433	160
10 - 19	22	377	2,019	4,381	324	10,148	..,516	89
20 - 39	57	648	3,308	6,384	217	9,6[illegible]	1,329	36
NA*	0	14	99	484	0	1,086	[illegible]94	0
40 - 59	0	98	1,081	2,485	216	5,010	413	0
60 - 79	0	152	1,125	2,362	9	2,648	45	0
80+	79	4,005	13,719	9,796	152	5,660	221	0

* Not available.

Table C-2. Aggregate measures of size for the Restandardization Study by cell for public schools in other school districts located in counties with 1960 population of 50,000 or more (major stratum groups and numbers are indicated)

Percent Minority	Average Adjusted Gross Income (in thousands)								
	<2	2 - 2.9	3 - 3.9	4 - 4.9	5 - 5.9	NA*	6 - 7.9	8 - 9.9	10+
<5	0	97	1,559	11,476	48,050	2,541	139,824	52,936	1,321
5 - 9	0	22	683	3,684	14,856	802	40,3??	12,586	83
10 - 19	0	0	737	4,759	17,473	957	33,637	10,063	409
20 - 39	4	14	1,775	7,214	20,144	616	27,062	6,491	58
40 - 59	0	31	656	3,166	7,109	352	7,958	567	
60 - 79	0	0	597	2,037	3,422	85	3,565	264	
80+	0	36	3,557	6,084	7,415	268	5,585	410	

* Not available.

Table C-3. Aggregate measures of size for the Restanderdization Study by cell for public schools in other school districts located in counties with 1960 population of less than 50,000 (major stratum groups and numbers are indicated)

Percent Minority	Median Family Income (in thousands)							
	<2	2 - 2.9	3 - 3.9	NA*	4 - 4.9	5 - 5.9	6 - 7.9	8 - 9.9
<5	2,818	15,087	28,756	12,352	49,575	31,089	2,778	72
5 - 9	457	3,313	5,856	1,588	6,631	4,018	950	14
10 - 19	625	4,317	7,727	1,300	5,255	2,994	702	51
20 - 39	1,337	7,933	10,792	775	5,046	1,945	349	14
40 - 59	1,351	7,831	4,901	291	2,675	926	167	14
60 - 79	1,927	3,680	1,625	297	803	313	41	
80+	1,766	2,784	1,284	1,381	1,180	515	65	
NA*				10,264				

* Not available.

Table C-4. Aggregate measures of size for the Restandardization Study by income class for Catholic schools (major stratum groups and numbers are indicated)

Median Income	Aggregate Measure of Size	
< $ 3,000	378	42
$ 3,000 - $ 4,999	7,928	
$ 5,000 - $ 6,999	9,055	
$ 7,000 - $ 8,999	47,191	43
$ 9,000 - $ 9,999	8,014	45
$10,000 - $11,999	3,109	
$12,000 - $13,999	4,994	46
$14,000 - $15,999	7,118	
$16,000 - $19,999	8,779	47
$20,000 +	1,043	
Not Available	13,137	44
Total	110,746	
Non-Catholic Schools	21,273	41

Table C-5. Aggregate measures of size for the Equating Study by cell for public schools in the 50 largest school districts (major stratum groups and numbers are indicated)

Percent Minority	Average Adjusted Gross Income (in thousands)									
	2 - 2.9	3 - 3.9	4 - 4.9	5 - 5.9	6 - 6.9	NA*	7 - 7.9	8 - 8.9	9 - 9.9	10+
<5	0	132	836	4,181	6,007	436	5,549	4,748	2,296	895
5 - 9	0	88	402	1,601	2,335	102	1,692	1,690	586	146
10 - 19	22	358	1,668	3,860	5,206	296	4,257	1,170	246	88
20 - 39	57	550	3,094	5,538	5,586	181	3,000	982	137	22
40 - 59	0	94	900	2,250	3,326	215	1,252	379	0	0
60 - 79	0	168	964	2,500	1,783	0	450	36	0	0
80+	79	3,604	11,776	8,653	3,980	88	847	181	44	0
NA*	0	24	101	442	312	0	728	168	0	0

24 19 18 7 11

* Not available.

Table C-6. Aggregate measures of size for the Equating Study by cell for public schools in other school districts located in counties with 1960 population of 50,000 or more (major stratum groups and numbers are indicated)

Percent Minority	Average adjusted gross income (in thousands)									
	2 - 2.9	3 - 3.9	4 - 4.9	5 - 5.9	NA*	6 - 6.9	7 - 7.9	8 - 8.9	9 - 9.9	10+
<5	88	1,023	8,281	37,060	2,126	63,524	51,519	31,847	12,770	988
5 - 9	31	419	2,805	11,880	546	18,894	14,669	8,571	2,205	57
10 - 19	0	564	3,618	14,187	883	16,275	11,854	7,293	1,178	391
20 - 39	22	1,244	6,163	16,150	334	13,089	8,713	4,315	898	66
40 - 59	22	491	2,436	5,793	253	4,255	2,044	372	14	0
60 - 79	0	412	1,566	2,592	44	2,227	677	159	58	0
80+	28	3,035	5,050	6,169	256	3,426	1,222	208	132	0

* Not available.

Table C-7. Aggregate measures of size for the Equating Study by cell for public schools in other school districts located in counties with 1960 population of less than 50,000 (major stratum groups and numbers are indicated)

Percent Minority	Median Family Income (in thousands)										
	<1	1 - 1.9	2 - 2.9	3 - 3.9	NA*	4 - 4.9	5 - 5.9	6 - 6.9	7 - 7.9	8 - 8.9	9 - 9.9
<5	68	1,925	10,791	19,419	9,357	34,692	22,563	1,906	9	0	22
5 - 9	106	248	2,603	4,409	1,277	4,266	2,949	672	0	0	0
10 - 19	343	194	3,300	5,902	920	3,903	2,069	510	0	0	44
20 - 39	194	718	6,155	8,125	681	3,985	1,509	373	0	0	14
40 - 59	127	1,037	6,494	3,679	250	2,122	733	123	0	22	0
60 - 79	93	1,469	2,613	1,275	248	647	267	14	0	0	0
80+	247	1,165	2,189	920	1,147	925	285	14	0	0	0
NA*	0	0	0	0	7,729	0	0	0	0	0	0

21 25 20 17 3 2

* Not available.

Table C-8. Aggregate measures of size for the Equating Study by income class for Catholic schools (major stratum groups and numbers are indicated)

Median income	Aggregate Measure of Size	
<$ 3,000	171	27
$ 3,000 - $ 4,999	5,538	
$ 5,000 - $ 6,999	7,031	
$ 7,000 - $ 8,999	36,276	22
Not Available	9,720	10
$ 9,000 - $ 9,999	6,434	
$10,000 - $11,999	2,703	
$12,000 - $13,999	4,136	
$14,000 - $15,999	6,004	4
$16,000 - $19,999	7,491	
$20,000 +	819	

APPENDIX D

APPROXIMATION OF DISTRICT SELECTION PROBABILITIES

As indicated in Section 7.2 of Chapter 3, the probability that at least one of the schools in a school district would be selected for the primary Restandardization sample was approximated for each district represented in the primary sample. If Q_d represents the probability that school district d would not be represented in the primary Restandardization sample, then the basic formula used to compute P_d is the following:

$$P_d = 1 - Q_d = 1 - Q_{1d}\, Q_{2d}\, Q_{3d} \cdots Q_{470d}\,, \qquad \text{(D-1)}$$

where

Q_{hd} = the probability that no schools from school district d would be selected from final stratum h.

Equation D-1 follows from the independence of selection from stratum to stratum.

Since the schools in each final stratum were ordered randomly prior to being sampled from the major stratum list (as described in Section 5.4 of Chapter 3), the probability, Q_{hd}, that district d would not be represented in the sample from final stratum h is given by the following expression:

$$Q_{hd} = \sum_{\text{all } j} P(\text{district d is not represented in stratum h given randomization j}) \times P(\text{randomization j}) \qquad \text{(D-2)}$$

A randomization refers to a particular random ordering of the schools in final stratum h.

Computing the probability, Q_{hd}, as specified in equation D-2, would have been a very tedious operation due to the large number of possible randomizations. Therefore, rather than carrying out the enumeration required by equation D-2, a much simpler method, described below, was used to approximate the Q_{hd} values.

Let A_h represent the aggregate measure of size for final stratum h (i.e., the total of the measures of size of all the schools in final stratum h). Also, let A_{hd} be the total of the measures of size of those schools in final stratum h that are contained in district d. Then, the proportion, Q'_{hd}, of the aggregate measure of size of final stratum h not accounted for by schools from district d is given as follows:

$$Q'_{hd} = (A_h - A_{hd})/A_h . \qquad \text{(D-3)}$$

Therefore, since two schools were selected from each final stratum with probabilities proportional to size, the probability that a school from district d would not be selected on a particular draw from final stratum h is Q'_{hd}, the proportion of the stratum aggregate measure of size not accounted for by district d. The probability, Q_{hd}, that district d would not be represented on both draws from final stratum h is given as follows:

$$Q_{hd} = P(\text{district d not represented on 1st draw}) \times P(\text{district d not represented on 2nd draw given it was not represented on 1st draw}). \qquad \text{(D-4)}$$

The first probability in the above expression is equal to Q'_{hd} as indicated above. The second probability, which would be very difficult to compute exactly, is only slightly less than Q'_{hd}. Therefore, this conditional probability was approximated by Q'_{hd}.

The resulting simple expression used to approximate Q_{hd} is the following:

$$Q_{hd} \doteq (Q'_{hd})^2 . \qquad \text{(D-5)}$$

Therefore, the district-selection probability given in equation D-1 was approximated by the following expression:

$$P_d \doteq 1 - (Q'_{1d}\, Q'_{2d} \cdots Q'_{470d})^2 . \qquad \text{(D-6)}$$

Of course, a majority of the 470 Q'_{hd} values were one for a particular district since most of the final strata did not contain schools from a particular district.

Since the expression used to approximate the Q_{hd} values given in equation D-5 above yielded slight overestimates of the Q_{hd} values, the resulting estimates of the P_d values obtained by using equation D-6 were slight underestimates. Since weights were computed basically as inverses of selection probabilities, the weights assigned to pupils participating in the supplementary sample were slightly higher than ideal. However, since the supplementary sample was quite small, the resulting bias from the procedure used to approximate the P_d values is trivial.

APPENDIX E

SURVEY ESTIMATES AND CORRESPONDING VARIANCE ESTIMATES FOR THE RESTANDARDIZATION STUDY

E.1 Introduction

As described in Section 5 of this report, the main portion of the Restandardization sample consisted of 240 schools, plus substitutes, selected from those listed on the data tapes used in the study. In addition to this, there was a procedure carried out to obtain a sample of those schools that were eligible for the study but omitted from the data tapes. (This sample supplementation procedure is described in Section 7 of Chapter 3.)

Though this supplemental sample was quite small,* its inclusion in the study caused complications in the variance estimation expressions. Consequently, to simplify the presentation to some extent, the first two estimates described in this appendix--a percentile rank and its precision--are discussed in two stages. First, the expressions for the component of each of these two estimates corresponding to the main portion of the sample (i.e., the primary sample plus substitutes) are given and denoted by the subscript "1." Next, the expressions for the component of each of these estimates corresponding to the supplemental sample are given and denoted by the subscript "2." Then the overall (or combined) estimates are presented.

---

* The supplemental sample of missed schools consisted of only 20 schools, of which 12 schools participated.

The following estimates are described in this appendix: (1) pupil percentiles and percentile ranks and their precision (Sections E.2, E.3, and E.4 and E.5), (2) the pupil mean and its precision (Section E.6), and (3) school percentiles and percentile ranks and their precision (Section E.7). Of these estimates, only pupil percentiles and percentile ranks have been computed.

E.2 Estimates of Pupil Percentile Ranks from the Main Portion of the Restandardization Sample

The percentile rank corresponding to a particular test score, T, is the proportion, P, of pupils in the population who have (expected) scores less than T. The P^{th} percentile for a test is that score, T_P, such that the proportion of all pupils who have (expected) scores less than T_P is P. That is, T_P is the P^{th} percentile if

$$P = X/Y,$$

where

Y = the total number of pupils in the population (in a particular grade), and

X = the number of pupils in the specific grade who have (expected) scores less than T.

Also, P is the percentile rank for the score T_P.

The sample estimate, p, of the percentile rank, P, corresponding to a particular score, T, was computed as follows:

$$p = x'/y' , \tag{1}$$

where

y' is the sample estimate of Y, and

x is the sample estimate of the number of pupils who have (expected) scores less than T.

The sample estimate of the P^{th} percentile, T_p, was taken to be that score, t_p, such that

$$P = x'/y' , \tag{2}$$

where

y' is the sample estimate of Y, and

x' is the sample estimate of the number of pupils who have (expected) scores less than t_p.

The estimated totals, x' and y', are composed of two components. That is, these estimates were computed using the following expressions:

$$x' = x'_1 + x'_2 , \text{ and}$$
$$y' = y'_1 + y'_2 , \tag{3}$$

where

x'_1 and y'_1 represent estimated totals based on the main portion of the sample, and

x'_2 and y'_2 represent estimated totals based on the supplementary sample of missed schools.

In this section, the expressions used to compute the main components, x'_1 and y'_1, of x' and y' are given. The corresponding expressions for x'_2 and y'_2 are given later in Section E.4.

The expression used to compute the main component, y'_1, of the estimate, y', is the following:

$$y'_1 = \sum_{h}^{L} \sum_{i}^{m_h} \sum_{k}^{n_{hi}} W_{hik} , \qquad (4)$$

where

L = 940 = the total number of strata;

m_h = the number of schools selected from final stratum h that cooperated in the study, including any substitutes;

n_{hi} = the number of usable test scores obtained from the i^{th} school selected from stratum h; and

W_{hik} = the sample weight, including adjustments for nonresponse, of the k^{th} participating pupil contained in the i^{th} school selected from stratum h.*

---

* The computation of sample weights is described in Section 9 of Chapter 3. The subscript "j" included in the weights in Section 9 to index the weighting class that contained a particular pupil has been omitted in this appendix to simplify the notation.

The expression used for the component of the estimate, x', corresponding to the primary sample is:

$$x'_1 = \sum_{h}^{L} \sum_{i}^{m_h} \sum_{k}^{n_{hi}} W_{hik}\, \delta_{hik}, \qquad (5)$$

where

$$\delta_{hik} = \begin{cases} 0 \text{ if the } hik^{th} \text{ score is greater than the particular score (T or } t_p). \\ 1 \text{ if the } hik^{th} \text{ score is less than the particular score (T or } t_p). \end{cases}$$

In the above estimates, x'_1 and y'_1, m_h took on the value 2 except in cases where one (or both) of the schools selected from a final stratum did not participate and a substitute school could not be obtained. Then, of course, m_h would be 1 (or 0).

E.3 Estimation of the Variances of Percentile Ranks from the Main Portion of the Restandardization Sample

As indicated in the previous section, the estimated percentile rank, p, corresponding to score T is a ratio estimate (x'/y'). The expression that can therefore be used to estimate the variance of a percentile rank, σ_p^2, is the following:*

$$\sigma_p^2 \doteq p^2 \left(\frac{\sigma_{x'}^2}{(x')^2} + \frac{\sigma_{y'}^2}{(y')^2} - \frac{2\sigma_{x'y'}}{x'y'} \right), \qquad (6)$$

---

* For a statement of the equation used to obtain the expression for V_p^2, see Hansen, Hurwitz, and Madow, Survey Sample Methods and Theory, Vol. I, p. 166, equation (18.12).

where

$\sigma^2_{x'}$ = the variance of x',

$\sigma^2_{y'}$ = the variance of y', and

$\sigma_{x'y'}$ = the covariance between x' and y'.

The expressions for the estimates $\sigma^2_{x'}$, $\sigma^2_{y'}$, and $\sigma_{x'y'}$ are all developed in two stages, as follows:

$$\begin{aligned} \sigma^2_{x'} &= \sigma^2_{x'_1} + \sigma^2_{x'_2}, \\ \sigma^2_{y'} &= \sigma^2_{y'_1} + \sigma^2_{y'_2}, \text{ and} \\ \sigma_{x'y'} &= \sigma_{x'_1y'_1} + \sigma_{x'_2y'_2}. \end{aligned} \tag{7}$$

As in the previous section, the components in the above equations with the subscript "1" denote those obtained from the main portion of the sample, while those components with the subscript "2" denote those obtained from the supplementary sample of missed schools. Sample estimates of $\sigma^2_{x'}$, $\sigma^2_{y'}$, and $\sigma_{x'y'}$ are obtained by computing sample estimates of each of the components in the above equations.

In this section, estimates of the components corresponding to the main portion of the sample are discussed. Estimates of the other components are discussed in the next section.

The sample estimate, $s^2_{x'_1}$, of $\sigma^2_{x'_1}$, should be computed using the following expression:

$$s^2_{x'_1} = \sum_h^{L'} (x'_{h1} - x'_{h2})^2 , \qquad (8)$$

where

L' = the number of final strata that contained two participating schools;

$$x'_{h1} = \sum_k^{n_{h1}} W_{h1k}\, \delta_{h1k} , \text{ and}$$

$$x'_{h2} = \sum_k^{n_{h2}} W_{h2k}\, \delta_{h2k} .$$

In the above equation, x'_{h1} and x'_{h2} are both estimates of half the total number of pupils in stratum h who have (expected) scores less than T.

The estimate, $\sigma^2_{x'_1}$, is computed only over those strata that contain two cooperating schools. Since the response rate was extremely high in this study, we do not recommend that the strata containing only one cooperating school be used in varaince estimation or that adjustments to the variance estimates be made to account for them. As a result, slight overestimates of variances may be obtained.

In a way analogous to the computation of $s^2_{x'_1}$, the main component, $s^2_{y'_1}$, of the sample estimate of $\sigma^2_{y'_1}$ should be computed as follows:

$$s^2_{y'_1} = \sum_h^{L'} (y'_{h1} - y'_{h2})^2 , \qquad (9)$$

where

$$y'_{hi} = \sum_{k}^{n_{hi}} W_{hik} , \quad i = 1, 2 .$$

Similarly, the main component, $s_{x'_1 y'_1}$, of the sample estimate of $\sigma_{x'y'}$ should be computed as follows:

$$s_{x'_1 y'_1} = \sum_{h}^{L'} (x'_{h1} - x'_{h2})(y'_{h1} - y'_{h2}) . \qquad (10)$$

E.4 Overall Estimates of Pupil Percentiles and Percentile Ranks

As discussed in detail in Section 7 of Chapter 3, the supplemental sample of missed schools was obtained by first stratifying all the public school districts represented in the primary sample by geographic region and size.* Next, two districts were selected from 78 of the 82 strata defined. All the districts contained in the other 4 strata were selected. In total, 178 districts were selected. From each of these 178 districts, a list of schools omitted from the public school file was obtained. Then a probability sample of missed schools was selected to be added to the study.

---

* The measure of size used was the estimated probability that at least one school in the district would be selected into the primary sample.

The estimate, y_2', of the total number of pupils in a specific grade contained in public schools omitted from the school file was computed as follows:

$$y_2' = \sum_{h}^{L_d} \sum_{j}^{m_h} \sum_{i}^{m_{hj}} \sum_{k}^{n_{hji}} W_{hjik}, \qquad (11)$$

where

L_d = 82 = the number of district strata,

m_h = the number of districts selected from district stratum h,

m_{hj} = the number of missed schools selected and participating from the j^{th} district selected from district stratum h,

n_{hji} = the number of usable test scores obtained from the hji^{th} school, and

W_{hjik} = the sample weight, including nonresponse adjustments, for the k^{th} participating student in the hji^{th} school.*

The expression above for y_2' is simply the sum of the weights of all participating pupils contained in the supplemental sample of missed schools.

---

* In this appendix, the subscript "j" is used to index the district. However, in the description of the weighting procedures given in Section 9 of Chapter 3, the subscript "j" denoted a particular weighting class.

In a similar way, the estimate, x'_2, of all pupils in missed schools who have (expected) scores less than score T was computed as follows:

$$x'_2 = \sum_h^{L_d} \sum_j^{m_h} \sum_i^{m_{hj}} \sum_k^{n_{hji}} W_{hjik}\, \delta_{hjik} , \qquad (12)$$

where

$$\delta_{hjik} = \begin{cases} 0 \text{ if the } hjik^{th} \text{ score is greater than T} \\ 1 \text{ if the } hjik^{th} \text{ score is less than T.} \end{cases}$$

As indicated previously, the components given in equations 11 and 12 above, along with those given earlier in Section E.2 in equations 4 and 5, were substituted into equation 3 to obtain estimates of x' and y'. From these estimates, the percentile rank for score T was computed as the ratio of x' to y'.

Also, based on the estimated totals, x' and y', an estimate of a particular percentile was made by finding that score having a particular estimated percentile rank. This estimation procedure was described in equation 2 in Section E.2.

E.5 Overall Estimates of the Variances of Percentiles and Percentile Ranks

The estimate, $s^2_{y'_2}$, of the variance of y'_2 given in equation 11 above should be computed as follows:

$$s^2_{y'_2} = \sum_h^{L_d} \frac{4}{m_h(m_h - 1)} \sum_j^{m_h} (y'_{hj} - \overline{y'_h})^2 , \qquad (13)$$

where

$$y'_{hj} = \sum_{i}^{n_{hj}} \sum_{k}^{r_{hji}} w_{hjik} \text{ , and}$$

$$\overline{y'_h} = (1/m_h) \sum_{j}^{m_h} y'_{hj} \text{ .}$$

In deriving the expression given in equation 13 above, it was assumed that all public school districts in the population were stratified and sampled rather than only those districts represented in the primary sample. Also, this expression was derived as though the subsampling of missed schools was carried out separately within each selected district. The estimate, $s^2_{y'_2}$, will probably be an overestimate of $\sigma^2_{y'_2}$, though this has not been determined analytically. It is unlikely that this overestimate will have more than a trivial effect on the overall variance estimates since the supplementary sample of missed schools was very small. It should be noted further that most of the y_{hj} values will be zero, either because the hj^{th} district reported no missed schools, or because the missed schools reported by the hj^{th} district either were not selected in the supplemental sample or were selected but refused to participate.

In a way similar to that for $s^2_{y'_2}$, the estimate, $s^2_{x'_2}$, of the compound of variance, $\sigma^2_{x'_2}$, should be computed as follows:

$$s^2_{x'_2} = \sum_{h}^{L_d} \frac{4}{m_h(m_h - 1)} \sum_{j}^{m_h} (x'_{hj} - \overline{x'_h})^2 \text{ ,} \qquad (14)$$

where

$$x'_{hj} = \sum_{i}^{m_{hj}} \sum_{k}^{n_{hji}} W_{hjik}\,\delta_{hjik}\,, \text{ and}$$

$$\overline{x'_h} = (1/m_h) \sum_{j}^{m_h} x'_{hj}\,.$$

Similarly, the estimate, $s_{x'_2y'_2}$, should be computed as follows:

$$s_{x'_2y'_2} = \sum_{h}^{L_d} \frac{4}{m_h(m_h - 1)} \sum_{j}^{m_h} (x'_{hj} - \overline{x'_h})(y'_{hj} - \overline{y'_h})\,. \qquad (15)$$

Using the estimates $s^2_{x'_2}$, $s^2_{y'_2}$, and $s_{x'_2y'_2}$ given above, along with the estimates $s^2_{x'_1}$, $s^2_{y'_1}$, and $s_{x'_1y'_1}$ given earlier in equations 8, 9, and 10, estimates of $\sigma_{x'}$, $\sigma_{y'}$, and $\sigma_{x'y'}$ can be computed as specified in equation 7 above. Then, the variance of an estimated percentile rank can be estimated by substituting the estimates from equation 7 into equation 6. An estimate of the standard error, σ_p, of an estimated percentile rank is obtained by simply taking the square root of the corresponding estimate of the variance, σ^2_p.

Estimates of the standard errors of percentiles are computed using the standard errors of percentile ranks as described in Appendix A. That is, for a particular percentile, T_p, a 68-percent confidence interval is first computed as the sample estimates, t_{P_U} and t_{P_L}, of the $P_U{}^{th}$ and $P_L{}^{th}$ percentiles, where

$$P_U = P + \hat{\sigma}_p\,, \text{ and}$$

$$P_L = P - \hat{\sigma}_p\,.$$

The term "$\hat{\sigma}_p$" represents, of course, the estimate of the standard error of p, computed as described above. From the two confidence limits, t_{P_U} and t_{P_L}, an estimate of the standard error, σ_{t_P}, of the estimated percentile is computed as follows:

$$\hat{\sigma}_{t_P} = \frac{1}{2}\left(t_{P_U} - t_{P_L}\right) . \qquad (16)$$

The estimates of σ_{t_P} will probably not be very reliable for values of P that are near zero or 1 (e.g., values of P less than .1 or greater than .9).

E.5 Estimating the Mean Score for All Pupils and the Standard Error of the Mean

The estimated mean score, $\bar{x}$, for a particular test is a ratio estimate similar to the estimate of the percentile rank, p, given in equation 1 in Section E.2. That is,

$$\bar{x} = x'/y' , \qquad (17)$$

where

y' is the estimated number of students as before,

$$x' = \sum_{h}^{L} \sum_{i}^{m_h} \sum_{k}^{n_{hi}} W_{hik}\, x_{hik} + \sum_{h}^{L_d} \sum_{j}^{m_h} \sum_{i}^{m_{hj}} \sum_{k}^{n_{hji}} W_{hjik}\, x_{hjik} ,$$

x_{hik} = the score on the test made by the k^{th} pupil participating from the i^{th} school in final stratum h in the main portion of the sample, and

x_{hjik} = the score on the test made by the k^{th} pupil participating from the i^{th} school in the j^{th} district selected from district stratum h in the sample supplement.

The estimate, x', is the estimated total of the (expected) scores on the test of all pupils in the population.

Because of the close similarity between the expression for $\bar{x}$ and that for p, the estimate, $\hat{\delta}_{\bar{x}}$, of the standard error of $\bar{x}$ is computed the same way $\hat{\delta}_p$ is computed, with δ_{hik} replaced by x_{hik} and δ_{hjik} replaced by x_{hjik}.

E.7 Estimates of School Percentiles and the Standard Errors of School Percentiles

Percentile ranks and percentiles for schools for a particular test are defined in a way similar to those for pupils. That is, T_P is the P^{th} percentile for schools for a given test if

$$P = X/Y ,$$

where

Y = the total number of schools in the population that contain the particular grade. and

X = the number of schools having mean (or median) scores less than T_P.

Correspondingly, P is the percentile rank for the score T_P.

The expressions for estimates of school percentiles and percentile ranks and their precision are very similar to those for pupils. Therefore, the equations for school estimates will be given in this section without any discussion of them.

First, the estimate of the total number of schools in the population containing a particular grade is

$$y' = \sum_{h}^{L} \sum_{i}^{m_h} W_{hi}\, \alpha_{hi} + \sum_{h}^{L_d} \sum_{j}^{m_h} \sum_{i}^{m_{hj}} W_{hji}\, \alpha_{hji}\,, \tag{18}$$

where

W_{hi} = the sample weight, including any adjustments for nonresponse, for the i^{th} participating school in (final) stratum h,

W_{hji} = the sample weight for the i^{th} school participating in the sample supplement, from the j^{th} district selected in district stratum h,

$$\alpha_{hi} = \begin{cases} 0 \text{ if the } hi^{th} \text{ school does not contain the particular grade} \\ 1 \text{ if the } hi^{th} \text{ school does contain the particular grade, and} \end{cases}$$

α_{hji} is defined for school hji like α_{hi} is defined for school hi.

The estimates of the total number of schools in the given grade that have mean (or median) scores less than a given score T is

$$x' = \sum_{h}^{L} \sum_{i}^{m_h} W_{hi}\, \delta_{hi} + \sum_{h}^{L_d} \sum_{j}^{m_h} \sum_{i}^{m_{hj}} W_{hji}\, \delta_{hji}\,, \tag{19}$$

where

$$\delta_{hi} = \begin{cases} 0 \text{ if the mean (or median) score}^{*} \text{ for school hi for the particular grade is greater than T, or if the school does not contain the particular grade} \\ 1 \text{ if the mean score for school hi for the particular grade is less than T, and} \end{cases}$$

δ_{hji} is defined for school hji like δ_{hi} is defined for school hi.

The estimate of the percentile rank (for schools) corresponding to a score T is

$$p = x'/y' . \tag{20}$$

As in the previous case, the p^{th} percentile is estimated by finding that score x_p whose estimated percentile rank is P.

The variance of x' is estimated by the following expression:

$$s^2_{x'} = s^2_{x'_1} + s^2_{x'_2} , \tag{21}$$

---

* The mean (or median) score for a particular school can be taken as a weighted or unweighted mean (or median) of the pupil test scores. Perhaps a weighted mean (or median) would be more appropriate since the pupil weights include nonresponse adjustments. The unweighted school mean (or median) would not contain any adjustments for pupil nonresponse.

where

$$s^2_{x'_1} = \sum_h^{L'} (W_{h1}\,\delta_{h1} - W_{h2}\,\delta_{h2})^2 ,$$

$$s^2_{x'_2} = \sum_h^{L_d} \frac{4}{m_h(m_h - 1)} \sum_j^{m_h} (x'_{hj} - \overline{x'_h})^2 ,$$

$$x'_{hj} = \sum_i^{m_{hj}} W_{hji}\,\delta_{hji} , \text{ and}$$

$$\overline{x'_h} = (1/m_h) \sum_j^{m_h} x'_{hj} .$$

The variance of y' is estimated similarly:

$$s^2_{y'} = s^2_{y'_1} + s^2_{y'_2} , \tag{22}$$

where

$$s^2_{y'_1} = \sum_h^{L'} (W_{h1}\,\alpha_{h1} - W_{h2}\,\alpha_{h2})^2 ,$$

$$s^2_{y'_2} = \sum_h^{L_d} \frac{4}{m_h(m_h - 1)} \sum_j^{m_h} (y'_{hj} - \overline{y'_h})^2 ,$$

$$y'_{hj} = \sum_i^{m_{hj}} W_{hji}\,\alpha_{hji} , \text{ and}$$

$$\overline{y'_h} = (1/m_h) \sum_j^{m_h} y'_{hj} .$$

The covariance between x' and y' is estimated as follows:

$$s_{x'y'} = s_{x'_1y'_1} + s_{x'_2y'_2} , \tag{23}$$

where

$$s_{x'_1y'_1} = \sum_{h}^{L'} (W_{h1}\,\delta_{h1} - W_{h2}\,\delta_{h2})(W_{h1}\,\alpha_{h1} - W_{h2}\,\alpha_{h2}) , \text{ and}$$

$$s_{x'_2y'_2} = \sum_{h}^{L_d} \frac{4}{m_h(m_h - 1)} \sum_{j}^{m_h} (x'_{hj} - \overline{x'_h})(y'_{hj} - \overline{y'_h}) .$$

From the estimates p, x', y', $s^2_{x'}$, $s^2_{y'}$, and $s_{x'y'}$ of P, X, Y, $\sigma^2_{x'}$, $\sigma^2_{y'}$, and $\sigma_{x'y'}$ for schools, estimates of the variances, and standard errors of percentile ranks are computed for schools in the same way they are for pupils, as specified by equations 6 and 7 of Section E.3. Also, confidence intervals for percentiles and estimates of the standard errors of percentile estimates can be computed from estimates of percentile rank standard errors, as described at the end of Section E.5.

APPENDIX F

Classification Procedure for "Weighting Classes" and Definitions of Eligibility and Usable Cases

Classification Procedure for Trivariate Layout

For each major stratum in the Restandardization Study, each eligible student in a participating school was classified in a trivariate (sex x race x IQ level) frequency distribution as either a usable or nonusable case. The student's sex was determined by his response on the answer sheet. Race and IQ level were determined by response on the "For Test Administrator Use Only" section of the answer sheet. The racial categories were: American Indian, Black, Oriental, Spanish-surnamed, and White and Other. The IQ categories were: below 75, 75-89, 90-110, 111-125, and above 125. For all three breakdowns, a category was provided for non-response.

Since, in the Equating Study, a separate adjustment (factor 3) was made for loss of entire classes, only students in participating classes of participating schools were entered in the trivariate distribution tables as usable or unusable cases. The categorization procedure was otherwise identical to that for the Restandardization Study.

Definitions

An eligible case was one which did not fall in any of the following categories.

(1) student declared to be mentally retarded or physically handicapped in the "For Test Administrator Use Only" section of the answer sheet.

(2) student declared to be extremely and chronically emotionally handicapped (i.e., generally untestable) on the Irregularity Report submitted by

the test administrator.

(3) student declared to have very little or no knowledge of the English language on the Irregularity Report submitted by the test administrator.

All students classified as eligible under the above definition were included in the data analyses, either as usable or unusable cases.

A usable case was defined as one which did not fall in any of the following categories.

(1) blank answer sheet, indicating absence.

(2) answer sheet with fewer than 5 response attempts in either the vocabulary or reading comprehension subtest.

(3) incomplete answer sheet, where the Irregularity Report submitted by the test administrator indicated that the student was removed from the testing situation.

(4) answer sheets judged by ETS staff to be invalid on the basis of information supplied by the test administrator on the Irregularity Report. This included instances of illness during testing, temporary physical impairment or emotional reaction, improper conditions of administration, misplacement of responses on answer sheet (where not correctable), etc.

The unusable cases were accounted for through the weighting process but contributed no actual test data to the analyses.

APPENDIX G

Tukey-Cureton Smoothing Procedure

The required input to the Tukey-Cureton 7-point rolling average smoothing procedure was the observed frequency distribution of a given test X, where X represents any one of the seven tests. The smoothing of this distribution was accomplished as follows.

1. The raw score range, for the unsmoothed distribution in single raw score points intervals of X, designated X_i where $i = 0, 1, 2, \ldots, \max(X)$, with frequency at the i^{th} interval designated f_i, was extended by 3 units at both the lower and upper limits. These intervals were designated -1, -2, -3 and max(X)+1, max(X)+2, and max(X)+3, respectively. Frequencies of zero were assigned to each of these "extended values."

2. Each f_i, $i = 0, 1, \ldots, \max(X)$, was smoothed to a new value f'_i by application of the following linear combination of observed frequencies:

$$f'_i = \frac{-2f_{i-3} + 3f_{i-2} + 6f_{i-1} + 7f_i + 6f_{i+1} + 3f_{i+2} - 2f_{i+3}}{21}$$

where f'_i was rounded to integer form and if negative, set equal to zero.

3. The sum of f'_i, designated N', was computed and each f'_i was then multiplied by $\frac{N}{N'}$ to restore the total frequency to its original size, N. The resulting set of adjusted smoothed frequencies rounded to integer form, f'_i, constituted the final values for the smoothed distribution from which the percentile ranks were derived.

APPENDIX B

Notation

In the following, the use of a subscript enclosed by parentheses indicates the (restricted) domain of reference of the subscripted symbol; where a symbol appears without such a definer, the domain may be taken to be the entire group for which the symbol is meaningful. Parentheses may be read as "in"; for example,

$M_{j(k)}$ may be read as "the mean of X_j in column k",

and refers to the mean computed for all students in column k who took X_j, or the jk^{th} cell mean of X_j.

M_j would be read as the mean of X_j over its entire meaningful range, i.e., for all students who took X_j or for row j.

The use of a minus sign subscript enclosed in parentheses indicates that the diagonal cell was excluded from the domain of reference. For example,

$M_{j(-)}$ indicates the mean of X_j over its entire range of applicability excluding the diagonal cell, i.e., over the six off-diagonal cells of row j.

General

j =	a given row of Table 4-4 or 4-5	j = 1, ..., 7
j' =	a given row of Table 4-4 or 4-5	j'= 1, ..., 7
k =	a given column of Table 4-4 or 4-5	k = 1, ..., 7
k' =	a given column of Table 4-4 or 4-5	k'= 1, ..., 7

X_j = the test common to the j^{th} row of Table 4-5 (i.e., a test administered as a first test)

X_k = the test common to the k^{th} column of Table 4-5 (i.e., a test administered as a second test)

$X_{ij(k)}$ = score of student i on test X_j in column k ($i=1, \ldots, N'_{j(k)}$)

$X_{ik(j)}$ = score of student i on test X_k in row j ($i=1, \ldots, N'_{k(j)}$)

$N'_{j(k)}$ = actual number of students who took test X_j in column k (i.e., in cell j,k)

$N'_{k(j)}$ = actual number of students who took test X_k in row j (i.e., in cell j,k)

$w_{ij(k)}$ = final operational weight assigned to the i^{th} student who took test X_j in column k

$w_{ik(j)}$ = final operational weight assigned to the i^{th} student who took test X_k in row j

$N_{j(k)}$ = weighted number of students who took X_j in column k

$N_{k(j)}$ = weighted number of students who took X_k in row j

$M_{j(k)}$ = mean of X_j in column k (i.e., in cell j,k)

$M_{k(j)}$ = mean of X_k in row j (i.e., in cell j,k)

$V_{j(k)}$ = variance of X_j in column k (i.e., in cell j,k)

$V_{k(j)}$ = variance of X_k in row j (i.e., in cell j,k)

$Cov(jk)$ = covariance of X_j and X_k in cell j,k

$S_{j(k)}$ = sum of scores (weighted) on X_j in column k (i.e., in cell j,k)

$S_{k(j)}$ = sum of scores (weighted) on X_k in row j (i.e., in cell j,k)

$SS_{j(k)}$ = sum of squares of scores (weighted) on X_j in column k (i.e., in cell j,k)

$SS_{k(j)}$ = sum of squares of scores (weighted) on X_k in row j (i.e., in cell j,k)

$SCP(jk)$ = sum of crossproducts, $X_{j(k)} X_{k(j)}$, (weighted) in cell j,k; $j \neq k$

Procedure 1

t = equating trial number

$t = 0$ indicates preliminary equating trial

$t = 1, \ldots, f$ indicates the successive iterations of the operational equating procedure

$2 \leq f \leq 4$ indicates the final operational trial

N_k = weighted number of students who took test X_k (i.e., in column k) including the lower diagonal cell

$N_{j(-)}$ = weighted number of students who took test X_j (i.e., in row j) excluding the diagonal cell

M_k = mean of X_k (a column mean)

V_k = variance of X_k (a column variance)

$\hat{M}_{j(k)}$ = estimated mean of X_j in column k (estimated for cell j,k from column k)

$\hat{V}_{j(k)}$ = estimated variance of X_j in column k (estimated for cell j,k from column k)

$_t\hat{M}_j$ = estimated mean of X_j produced by the t^{th} equating trial

$_t\hat{V}_j$ = estimated variance of X_j produced by the $t^{\underline{th}}$ equating trial

$_tA_j$ = scale parameter obtained from $t^{\underline{th}}$ trial for linear equating of X_j to MAT

$_tB_j$ = location parameter obtained from trial t for linear equating of X_j to MAT

$\{_tX_j^*\}$ = MAT-equivalent scores (denoted by *), obtained from the $t^{\underline{th}}$ equating trial, for the set of X_j

$_tX_k^*$ = MAT-equivalent scores for X_k obtained by application of the set of conversion tables $\{_tX_j^*\}$ where $j = k$

$_tA'_k \equiv {}_tA_j$ = scale parameter obtained from trial t for converting X_k to MAT using parameters determined for converting X_j to MAT

$_tB'_k \equiv {}_tB_j$ = location parameter obtained from trial t for converting X_k to MAT using parameters determined for converting X_j to MAT

$M_{j(-)}$ = row mean of X_j, excluding the diagonal cell

$V_{j(-)}$ = row variance of X_j, excluding the diagonal cell

$_tM^*_{(j-)}$ = MAT-equivalent mean in the $j^{\underline{th}}$ row, excluding the diagonal cell, obtained by the $t^{\underline{th}}$ equating trial

$_tV^*_{(j-)}$ = MAT-equivalent variance in the $j^{\underline{th}}$ row, excluding the diagonal, obtained by the $t^{\underline{th}}$ equating trial

$_tCov(j*)_{(j-)}$ = covariance of X_j and $_tX_k^*$ in $j^{\underline{th}}$ row, excluding the diagonal

$_tM^*$ = mean of the MAT-equivalent scores obtained from trial t for the entire table

$_tV^*$ = variance of the MAT-equivalent scores obtained from trial t for the entire table

$A_{j \to j'}$ = scale parameter obtained from final variances (Procedure 3) or their estimates (Procedure 1) for equating X_j to $X_{j'}$

$B_{j \to j'}$ = location parameter obtained from final means and variances (Procedure 3) or their estimates (Procedure 1) for equating X_j to $X_{j'}$

$X_{j \to j'}$ = equivalent score determined by conversion of X_j to $X_{j'}$

Procedure 2

Notation is analogous to that for Procedure 1, interchanging j and k subscripts and references to rows and columns and to upper and lower diagonal cells accordingly.

Average of Procedures 1 and 2

$X_{\ell j \to j'}$ = equivalent score (from Procedure 1) on X_j, for ℓ^{th} raw score value of X_j; $\ell = 0, 1, \ldots, \max(X_j)$

$X_{\ell k \to k'}$ = equivalent score (from Procedure 2) on $X_{k'}$ for ℓ^{th} raw score value of X_k; $\ell = 0, 1, \ldots, \max(X_k)$

$\overline{X}_{\ell j \to j'}$ = average of equivalent scores, $X_{\ell j \to j'}$ and $X_{\ell k \to k'}$, where $j = k$ and $j' = k'$

Procedure 3

N_j = weighted number of students who took test X_j (i.e., in row j including the upper diagonal cell)

$\overline{M}_j$ = mean of X_j (for all students in row j) combined with X_k (for all students in column k), where $j = k$

$\overline{V}_j$ = pooled variance of X_j and X_k, where $j = k$

Estimation of Equating Error

r = replication number for half-sample estimation of equating error, $r = 1, \ldots, 8$

$\hat{E}(X^{*}_{\ell j})$ = estimated error in equating X_j to the MAT at ℓ^{th} raw score value

Conditional Error of Equating

$X_{i\ell_j}$ = raw score (ℓ) on test X_j for student i; $\ell = 0, 1, \ldots, \max(X_j)$

$X_{imj'}$ = raw score (m) on test $X_{j'}$ for student i; $m = 0, 1, \ldots, \max(X_{j'})$

$X_{i\ell,j\to j'}$ = converted score on $X_{j'}$ for $X_{\ell j}$ for student i

$CE(X_{\ell j\to j'})$ = conditional error of equating ℓ^{th} raw score on X_j to $X_{j'}$

n_ℓ = weighted number of students with ℓ^{th} raw score on X_j

$CE(X_{j\to j'})$ = conditional error of equating X_j to $X_{j'}$ over all raw score values $X_{\ell j}$

Correlations

r_{jk} = correlation between X_j and X_k $\quad j \neq k$

$r_{j\dagger}$ = correlation between X_j administered as a first test and its secondary form $\dagger$ administered as a second test; an alternate forms reliability coefficient

$r_{\dagger k}$ = correlation between the secondary form $\dagger$ administered first and the primary form X_k administered as a second test; an alternate forms reliability coefficient.

z_{jk} = Fisher's z-equivalent for r_{jk}

$\bar{z}_{jk}$ = average of Fisher's z-equivalents for r_{jk} and r_{kj}

$\bar{r}_{jk}$ = r-equivalent for $\bar{z}_{jk}$; correlation between X_j and X_k averaged over orders of administration

$\bar{r}_{jj}$ = alternate forms reliability of X_j averaged over orders of administration

$_c r_{jk}$ = correlation between X_j and X_k, r_{jk}, corrected for attenuation due to unreliability

$_c \bar{r}_{jk}$ = averaged correlation between X_j and X_k, $\bar{r}_{jk}$, corrected for attenuation due to unreliability

APPENDIX J

Frequency Estimation Procedure

The purpose of the frequency estimation procedure was to estimate a marginal frequency distribution for a test Y, given only to a smaller group, using both the bivariate distribution of Y with another test X given to the same small group and the univariate distribution of X given to a larger group. This procedure assumes the same regression system of Y on X for smaller and larger groups.

In the following, X and Y each successively designate, for each analogue of the general design, the seven tests for which distributions were estimated.

For the preliminary equating, the smaller group was the students in a single off-diagonal cell of Table 4-5 while the larger group was the students in the column containing that cell for Procedure 1 (or the row containing that cell for Procedure 2). For all operational iterations, the smaller group was the students in a single row of Table 4-5 for Procedure 1 (or a single column for Procedure 2) while the larger group was formed by the union of all students in the off-diagonal cells of Table 4-5 with those in the lower diagonal cells (for Procedure 1) or in the upper diagonal cells (for Procedure 2).

The estimation procedure required that, for each raw score value of X in the smaller group, the observed array distribution of Y be expanded in proportion to the frequency of that particular X value in the larger group. This was accomplished by multiplying each Y frequency in the array for a given X by the ratio of the frequency of the given X in the larger group to the frequency of the given X in the smaller group. Thus, the

frequency estimation procedure consisted of determining an appropriate ratio for each raw score value of X, use of these ratios to develop an estimated bivariate distribution of X and Y for the larger group, and use of the estimated bivariate distribution to determine the estimated marginal distribution of Y for the larger group.

Because, in the smaller group, frequencies at some score points of X might be too small to provide a reasonable estimate of the array distribution or might be disproportionate with respect to the corresponding frequencies of X in the larger group, some collapsing of arrays corresponding to adjacent score points was provided for. Two criteria were involved in the determination of whether or not such collapsing would take place.

(a) Primary Criterion: the ratio of the frequency in the larger group to the frequency in the smaller group for a given value of X must be less than R_{max}. If the primary criterion was satisfied, no collapsing was required. If not, the secondary criterion was invoked.

(b) Secondary Criterion: the frequency of a given value of X in the smaller group must be greater than K. If the secondary criterion was satisfied, no collapsing was required. If not, then adjacent score point arrays were collapsed until either the primary or secondary criterion was satisfied.

These criteria were applied successively, beginning with the lowest score and going to the median (the "upward" phase) and then beginning with the highest score and going to the median (the "downward" phase). The detailed procedure for application of the above criteria follows.

Notation

Let $[X_{it}]$ be the $i^{\underline{th}}$ interval at trial t. On the first trial for a given interval, $[X_{i1}]$ contained only a single raw score value X_ℓ. If collapsing was required, the second trial interval $[X_{i2}]$ contained both X_ℓ and $X_{\ell+1}$ if $[X_{it}]$ was being determined in the "upward" phase and both X_ℓ and $X_{\ell-1}$ if in the "downward" phase. If further collapsing was required, $[X_{i3}]$ contained the three scores X_ℓ, $X_{\ell+1}$, and $X_{\ell+2}$ ("upward"), or X_ℓ, $X_{\ell-1}$, and $X_{\ell-2}$ ("downward"), etc., until either the primary or secondary criterion was satisfied. At this point the final interval was designated $[X_{if}]$.

Let $F_w [X_{it}]$ designate the weighted frequency for the larger group and $f_w [X_{it}]$ and $f [X_{it}]$ designate respectively the weighted and unweighted frequencies for the smaller group, which accommodate the intervals so defined.

Application of Criteria

(1) The primary criterion applied to the $[X_{it}]$ is as follows:

Form the ratio: $R[X_{it}] = \dfrac{F_w [X_{it}]}{f_w [X_{it}]}$ 4 decimal accuracy

and test to see if: $R [X_{it}] \leq R_{max}$

(a) If satisfied, designate $[X_{it}]$ as the final interval $[X_{if}]$ and $R [X_{it}]$ as the final ratio $R [X_{if}]$. Then go to (4).

(b) If not satisfied, go to (2).

(2) Apply the secondary criterion to the unweighted frequency of $[X_{it}]$. Test to see if:

$$f [X_{it}] \geq K$$

(a) If satisfied, designate $[X_{it}]$ as the final interval $[X_{if}]$ and $R[X_{it}]$ as the final ratio $R[X_{if}]$. Then go to (4).

(b) If not satisfied, go to (3).

(3) Revise the interval $[X_{it}]$ to include the next adjacent score point (just above for the "upward" phase; just below for the "downward" phase), producing a new interval $[X_{i,t+1}]$. Determine the corresponding frequencies, $F_w[X_{i,t+1}]$, $f_w[X_{i,t+1}]$, and $f[X_{i,t+1}]$, and subject the new interval to criterion (1, and, if necessary, criterion (2).

(4) Assign the final $R[X_{if}]$ to each raw score X_ℓ contained in the final interval $[X_{if}]$. The ratio assigned to each raw score value X_ℓ is designated $R(X_\ell)$.

The operational procedure for order of interval formation and subsequent estimation of the marginal distribution of Y was as follows.

(A) A weighted single raw score points frequency distribution of X in the larger group was developed. That is, the weighted frequency $F_w(X_\ell)$ for each raw score point X_ℓ was determined.

(B) A weighted bivariate frequency distribution in single raw score point intervals of X and Y in the smaller group, with Y the ordinate and X the abscissa, was formed.

(C) Both the weighted and unweighted marginal frequency distributions for X in the smaller group were produced. That is, the weighted and unweighted frequencies, $f_w(X_\ell)$ and $f(X_\ell)$, respectively, for each raw score point X_ℓ were determined.

(D) For the first trial of the first interval in the "upward" phase, $[X_{11}]$ consisted of the lowest single raw score point X_ℓ for the larger group. This interval was tested via criterion (1) and, if necessary, criterion (2).

(E) For the first trial of the second interval in the "upward" phase, $[X_{21}]$ consisted of the raw score just above the upper bound of interval $[X_{1f}]$. This interval was tested via criterion (1) and, if necessary, criterion (2).

(F) Step E was repeated, incrementing the interval number (i) by 1, for each raw score value just above the upper bound of the previous interval until the rounded median score for the smaller group, X_m, was reached but not exceeded. (X_m=Med(X)+.5).

That is, the "upward" phase terminated at X_m if the interval for which X_m was the upper bound satisfied either criterion (1) or criterion (2). If this interval did not satisfy the criterion, then the "upward" phase terminated instead with the upper bound of the preceding interval, designated $X_{m'}$. Let X_{m*} be the terminus of the "upward" phase (i.e., either X_m or $X_{m'}$).

(G) For the first trial of the first interval in the "downward" phase, $[X_{11}]$ consisted of the highest single raw score point X_ℓ observed in the larger group. This interval was tested via criterion (1) and, if necessary, criterion (2).

(H) For the first trial of the second interval in the "downward" phase, $[X_{21}]$ consisted of the raw score just below the lower bound of $[X_{1f}]$. This interval was tested via criterion (1) and, if necessary, criterion (2).

(I) Step H was repeated, incrementing the interval number (i) by 1, for each raw score value just below the lower bound of the preceding interval until reaching the interval for which $X_{m*} + 1$ was the lower bound. If this interval satisfied either criterion (1) or criterion (2), the "downward" phase terminated. If this

interval did not satisfy the criterion, it was combined with the previous interval established by the "downward" phase, and the resulting interval was accepted as final, whether or not it satisfied either criterion, thus terminating the "downward" phase.

(J) Using the final set of ratios corresponding to each raw score value of X developed as a result of application of the preceding steps, each weighted frequency in the array corresponding to a given X_g in the weighted bivariate distribution from Step B was multiplied by $R(X_g)$.

(K) The application of Step J to each array of frequencies in the weighted bivariate distribution for the smaller group resulted in an estimated weighted bivariate distribution for the larger group. Hence, to get an estimated weighted marginal frequency for each raw score value of Y, all estimated frequencies in the row corresponding to that raw score value of Y (across all raw score values of X) were summed. The set of raw score values of Y, each paired with the frequency so obtained, was the estimated weighted marginal frequency distribution of Y.

APPENDIX K

Equipercentile Equating Procedure

The required input for application of the equipercentile equating procedure was the cumulative frequency distributions, derived from the smoothed frequency distributions, for both X, the test equated, and Y, the test to which X was equated. In the following, X and Y each successively designate, for each analogue of the general design, the seven tests in the equating scheme. (Since no test was equated to itself, X and Y did not simultaneously represent the same test.)

The details of the equipercentile equating procedure are described below.

1. The percentile rank (cumulative percentage) was computed for each observed raw score value X_ℓ, where X_ℓ represents the condition that X = ℓ^{th} raw score value of X; $\ell = 0, 1, \ldots, \max(X)$. This percentile rank was designated P_{X_ℓ}.

2. For each percentile rank, P_j, where P_j represents the condition that P = j and j = successively one of the following 31 percentile ranks (not necessarily observed): 0.0, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0, ... continuing by 5's ..., 90.0, 95.0, 96.0, 97.0, 98.0, 99.0, 99.5, 100.0, the corresponding raw score value (percentile), X_{P_j}, was determined by linear interpolation.

   (a) The smallest interval of observed P_{X_ℓ} such that

   $$P_{X_{\ell-1}} \leq P_j \leq P_{X_\ell}$$

   was determined.

   (b) Then $$X_{P_j} = X_{\ell-1} + .5 + \frac{P_j - P_{X_{\ell-1}}}{P_{X_\ell} - P_{X_{\ell-1}}}$$

(c) Special cases:

X_{P_0} = minimum observed score −.5 in the smoothed distribution

$X_{P_{100}}$ = maximum observed score +.5 in the smoothed distribution

This interpolation procedure yielded the set, designated $\{X_{P_j}\}$, of 31 ordered scores X_{P_j}, one corresponding to each of the 31 percentile ranks P_j.

3. The set $\{Y_{P_j}\}$ was determined by analogous application of Steps 1 and 2 to Y.

4. Each of the 31 ordered values in the set $\{X_{P_j}\}$ was paired with its corresponding value (i.e., for the same j) in the set $\{Y_{P_j}\}$.

5. For each integral raw score value X_ℓ such that

$$X_{P_0} \leq X_\ell \leq X_{P_{100}},$$

the smallest interval in the set $\{X_{P_j}\}$ such that

$$X_{P_{j-1}} \leq X_\ell \leq X_{P_j}$$

was determined, and the corresponding Y_{X_ℓ} was found by linear interpolation.

$$Y_{X_\ell} = Y_{P_{j-1}} + \left(\frac{X_\ell - X_{P_{j-1}}}{X_{P_j} - X_{P_{j-1}}} \right) \left(Y_{P_j} - Y_{P_{j-1}} \right)$$

6. For raw score values outside the range defined above, i.e., for $X_\ell > X_{P_{100}}$ and $X_\ell < X_{P_0}$, equivalent scores Y_{X_ℓ} were found by linear extrapolation using the two points $(X_{P_{100}}, Y_{P_{100}})$ and $(X_{P_{99.5}}, Y_{P_{99.5}})$ in the former case and the two points (X_{P_0}, Y_{P_0}) and $(X_{P_{0.5}}, Y_{P_{0.5}})$ in the latter.

(a) For $X_\ell > X_{P_{100}}$:

$$Y_{X_\ell} = \left(\frac{Y_{P_{100}} - Y_{P_{99.5}}}{X_{P_{100}} - X_{P_{99.5}}} \right) \left(X_\ell - X_{P_{100}} \right) + Y_{P_{100}}$$

If Y_{X_ℓ} determined by the above formula was greater than the maximum raw score value of Y, Y_{X_ℓ} was set equal to maximum raw score of Y; otherwise Y_{X_ℓ} was accepted as determined by the extrapolation.

(b) For $X_\ell < X_{P_0}$:

$$Y_{X_\ell} = \left(\frac{Y_{P_{0.5}} - Y_{P_0}}{X_{P_{0.5}} - X_{P_0}} \right) \left(X_\ell - X_{P_0} \right) + Y_{P_0}$$

If Y_{X_ℓ} determined by the above formula was less than zero, Y_{X_ℓ} was set equal to zero; otherwise, Y_{X_ℓ} was accepted as determined by extrapolation.

APPENDIX L

Computational Formulas

Linear Equating

Preliminary

$$N_{j(k)} = \sum_{i=1}^{N'_{j(k)}} w_{ij(k)}$$

$$N_{k(j)} = \sum_{i=1}^{N'_{k(j)}} w_{ik(j)}$$

$$N_{k} = \sum_{j=1}^{7} N_{k(j)}$$

$$N_{j} = \sum_{k=1}^{7} N_{j(k)}$$

$$N_{j(-)} = \sum_{\substack{k=1 \\ k \neq j}}^{7} N_{j(k)}$$

$$S_{j(k)} = \sum_{i=1}^{N'_{j(k)}} w_{ij(k)} X_{ij(k)}$$

$$S_{k(j)} = \sum_{i=1}^{N'_{k(j)}} w_{ik(j)} X_{ik(j)}$$

$$SS_{j(k)} = \sum_{i=1}^{N'_{j(k)}} w_{ij(k)} X^2_{ij(k)}$$

$$SS_{k(j)} = \sum_{i=1}^{N'_{k(j)}} w_{ik(j)} X^2_{ik(j)}$$

$$SCP_{(jk)} = \sum_{i=1}^{N'_{j(k)}} w_{ij(k)} X_{ij(k)} X_{ik(j)}$$

$$M_{j(-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} S_{j(k)}}{N_{j(-)}}$$

$$V_{j(-)} = \frac{N_{j(-)} \left[\sum_{\substack{k=1 \\ k\neq j}}^{7} SS_{j(k)} \right] - \left[\sum_{\substack{k=1 \\ k\neq j}}^{7} S_{j(k)} \right]^2}{N_{j(-)} \left[N_{j(-)} - 1 \right]}$$

Procedure 1

(1)

$$\hat{M}_{j(k)} = \frac{S_{j(k)}}{N_{j(k)}} + \left[\frac{N_{k(j)} - 1}{N_{k(j)}} \right] \left[\frac{N_{k(j)} \, SCP_{(jk)} - S_{j(k)} \, S_{k(j)}}{N_{k(j)} \, SS_{k(j)} - S^2_{k(j)}} \right] \left[\frac{\sum_{j=1}^{7} S_{k(j)}}{N_k} - \frac{S_{k(j)}}{N_{k(j)}} \right]$$

(2)

$$\hat{V}_{j(k)} = \frac{N_{j(k)} \, SS_{j(k)} - S^2_{j(k)}}{N_{j(k)} \, (N_{j(k)} - 1)} + \left[\frac{N_{k(j)} \, SCP_{(jk)} - S_{j(k)} \, S_{k(j)}}{N_{k(j)} \, SS_{k(j)} - S^2_{k(j)}} \right]^2 \cdot \left[\frac{N_{k(j)} - 1}{N_{k(j)}} \right]^2 \cdot \left[\frac{N_k \sum_{j=1}^{7} SS_{k(j)} - \left(\sum_{j=1}^{7} S_{k(j)} \right)^2}{N_k \, (N_k - 1)} \right] - \frac{N_{k(j)} \, SS_{k(j)} - S^2_{k(j)}}{N_{k(j)} \, (N_{k(j)} - 1)}$$

(3) $$ {}_0\hat{M}_j = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} N_k \, \hat{M}_{j(k)}}{\sum_{\substack{k=1 \\ k\neq j}}^{7} N_k} $$

(4) $$ {}_0\hat{V}_j = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} (N_k - 1) \, \hat{V}_{j(k)} + \sum_{\substack{k=1 \\ k\neq j}}^{7} N_k \left(\hat{M}_{j(k)} - {}_0\hat{M}_j \right)^2}{\left(\sum_{\substack{k=1 \\ k\neq j}}^{7} N_k \right) - 1} $$

(5) $$ {}_0A_j = \left(\frac{{}_0\hat{V}_4}{{}_0\hat{V}_j} \right)^{1/2} $$

(6) $$ {}_0B_j = {}_0\hat{M}_4 - {}_0A_j \, {}_0\hat{M}_j $$

(7) $$ {}_0X^*_j = {}_0A_j \, X_j + {}_0B_j $$

(8) $$ {}_0M^*_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} \left({}_0A'_k \, S_{k(j)} + N_{k(j)} \, {}_0B'_k \right)}{N_{j(-)}} $$

(9)

$$ {}_0V^*_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} \left[\frac{{}_0A'^2_k}{N_{k(j)}}\right]\left[N_{k(j)}\ SS_{k(j)} - S^2_{k(j)}\right] + \sum_{\substack{k=1 \\ k\neq j}}^{7} N_{k(j)}\left[{}_0A'_k \frac{S_{k(j)}}{N_{k(j)}} + {}_0B'_k - {}_0M^*_{(j-)}\right]^2}{N_{j(-)} - 1} $$

(10)

$$ {}_0Cov(j^*)_{(j-)} = \frac{\sum_{\substack{k=1 \\ k\neq j}}^{7} \left[{}_0A'_k\ SCP_{(jk)} + {}_0B'_k\ S_{j(k)}\right] - \left[\sum_{\substack{k=1 \\ k\neq j}}^{7} S_{j(k)}\right]\left[{}_0M^*_{(j-)}\right]}{N_{j(-)}} $$

(11)

$$ {}_0M^* = \frac{\sum_{k=1}^{7}\sum_{j=1}^{7}\left[{}_0A'_k\ S_{k(j)} + {}_0B'_k\ N_{k(j)}\right]}{\sum_{k=1}^{7}\sum_{j=1}^{7} N_{k(j)}} $$

(12)

$$ {}_0V^* = \frac{\sum_{k=1}^{7}\sum_{j=1}^{7}\left[\frac{{}_0A'^2_k}{N_{k(j)}}\right]\left[N_{k(j)}\ SS_{k(j)} - S^2_{k(j)}\right] + \sum_{k=1}^{7}\sum_{j=1}^{7} N_{k(j)}\left[{}_0A'_k \frac{S_{k(j)}}{N_{k(j)}} + {}_0B'_k - {}_0M^*\right]^2}{\left[\sum_{k=1}^{7}\sum_{j=1}^{7} N_{k(j)}\right] - 1} $$

(13)

$$ {}_1\hat{M}_j = M_{j(-)} + \left[\frac{{}_0Cov(j^*)_{(j-)}}{{}_0V^*_{(j-)}}\right]\left[{}_0M^* - {}_0M^*_{(j-)}\right] $$

(14)

$$ {}_1\hat{V}_j = V_{j(-)} + \left[\frac{{}_0Cov(j^*)_{(j-)}}{{}_0V^*_{(j-)}}\right]^2\left[{}_0V^* - {}_0V^*_{(j-)}\right] $$

(15) $$ {}_1A_j = \left(\frac{{}_1\hat{V}_4}{{}_1\hat{V}_j} \right)^{1/2} $$

(16) $$ {}_1B_j = {}_1\hat{M}_4 - {}_1A_j \; {}_1\hat{M}_j $$

(17) $$ {}_1X^*_j = {}_1A_j \; X_j + {}_1B_j $$

(18) $$ A_{j \to j'} = \left(\frac{{}_f\hat{V}_{j'}}{{}_f\hat{V}_j} \right)^{1/2} $$

(19) $$ B_{j \to j'} = {}_f\hat{M}_{j'} - A_{j \to j'} \; {}_f\hat{M}_j $$

(20) $$ X_{j \to j'} = A_{j \to j'} \; X_j + B_{j \to j'} $$

Procedure 2

Computational formulas are identical to those for Procedure 1, except that all j and k subscripts and indices of summation as well as references to rows and columns and to upper and lower diagonal cells are interchanged.

Average of Procedures 1 and 2

$$\bar{X}_{\ell j \to j'} = \tfrac{1}{2}\left(X_{\ell j \to j'} + X_{\ell k \to k'}\right) \qquad \begin{array}{l} j = k \\ j' = k' \\ \\ \text{for all } \ell \\ \text{for all } j' \end{array}$$

Procedure 3

$$\bar{M}_j = \frac{\sum_{k=1}^{\prime} S_{j(k)} + \sum_{j=1}^{7} S_{k(j)}}{N_k + N_j}$$

where j = k when not parenthesized

$$\bar{V}_j = \frac{(N_k + N_j)\left[\sum_{k=1}^{7} SS_{j(k)} + \sum_{j=1}^{7} SS_{k(j)}\right] - \left[\sum_{k=1}^{7} S_{j(k)} + \sum_{j=1}^{7} S_{k(j)}\right]^2}{(N_k + N_j)\,(N_k + N_j - 1)}$$

where j = k when not parenthesized

$$X_{j \to j'} = \left(\frac{\bar{V}_{j'}}{\bar{V}_j}\right)^{\frac{1}{2}} (X_j - \bar{M}_j) + \bar{M}_{j'}$$

Correlations

$$r_{jk} = \frac{\text{Cov } (jk)}{\sqrt{V_{j(k)} \; V_{k(j)}}}$$

$$z_{jk} = \tfrac{1}{2} \ln \left(\frac{1 + r_{jk}}{1 - r_{jk}} \right)$$

$$\bar{z}_{jk} = \frac{N_{j(k)} \; z_{jk} + N_{k(j)} \; z_{k}}{N_{j(k)} + N_{k(j)}}$$

$$\bar{r}_{jk} = \frac{e^{2\bar{z}_{jk}} - 1}{e^{2\bar{z}_{jk}} + 1}$$

$$c r_{jk} = \frac{r_{jk}}{\sqrt{\bar{r}_{j\dagger} \; \bar{r}_{\dagger k}}}$$